MODELLING NEWS PROGRESSION USING ENTITY-TIMELINES

Michael Stisen, Martin H. Thygesen

Aalborg University, Aalborg, Denmark

1. INTRODUCTION

With the massive amount of news available each day, it can be hard to gain a complete overview over what events are happening and who plays a part in the events. This information overload has two aspects:

Firstly it can be hard to figure out what events are important on a day to day basis due to the sheer volume of information presented. Solving this would require some measure of relevance to the reader for each piece of news, eg. if an article describing a drop in stock prices and an article describing politicians voting on whether or not they should allow fracking in an area are both in todays paper, one would have to be able to figure out which of those news would be most relevant to the user and show them only the relevant article.

Secondly, it can be hard to discern what the articles are portraying if news are just archived chronologically by publication date. This issue is more about how the information is presented, eg. if a law has been passed which allows for fracking in an area, it raises questions such as how long it has been discussed, who was involved in the voting, who voted what, whether any lobbyists were involved, what arguments were for and against and so on. A system solving this issue will have to grant the user a good overview of the events described in the news. We focus on the second problem of properly conveying information about who and what are involved in events throughout time based on archived news.

Providing an overview of archived news is a well studied area, although many focus on presenting topics [1–3]. Topic is a term covering several smaller events e.g. a topic such as "Aalborg University collaborates with city council" could have the smaller events "The city council set aside money for building projects on university campus" and "Architecture students to draw proposals for new sports center on campus". These approaches often group the events into sequences that represents a coherent story related to that issue. They do not however present a story of specific agents involved in the events. We are interested in conveying the story of these agents appearing in the news, e.g. we want to tell the story of how Barack Obama has appeared throughout the news during a span of time, rather than what happened during presidential inaugurations. To do this, we create timelines from the news articles based on the named entities, which are found as in [4], and can roughly be thought of as the proper nouns found in the text.

To represent information in a news archive, we use the concept of metro maps presented in [2], which represents several different timelines over the news articles and their intersections with each other. As our main purpose is to convey how named entities are involved in stories, we use metro lines to represent each named entity and metro stops to represent a cluster of news stories forming an event that the named entities take part in. If a named entity takes part in an event, its metro line intersects the metro stop representing that event. An example of this can be seen in figure 1.

In the example of a metro line in figure 1, we have 3 metro lines, each representing a named entity: The green line (1) representing the rector of Aalborg University, Per Michael Johansen , the blue line (2) representing the Department of Computer Science on Aalborg University and the red line (3) representing Aalborg University. Those 3 agents are part of a set of news stories covering 4 events, represented in figure 1 by the text in the intersection of the metro lines (4). Those events are made by finding representative sentences in the articles mentioning them based on the clustering done to find them. Lastly, in figure 1 the events are associated with a time (5) representing roughly when it takes place.

We concern ourselves mostly with small datasets from one or few sources. The dataset we use for testing is a set of newspaper articles provided by the Danish news agency NORDJYSKE. Although our approach should also be able to handle a multitude of sources each describing the same named entities and events, which many news archives may have, we focus on being able to accurately represent the story of less represented agents such as a local politician, who may be mentioned a few times a month from a local news source rather than representing an agent such as the U.S. president, who is often mentioned in several stories and events on a daily basis.

In the following we present an overview of the sections in the paper.

Section 2 contains preliminaries, under which we describe terminology and related work.

Section 3 describes the concept of named entities, what they are and how we obtain them.

Section 4 details how a metro map is constructed.

Section 5 gives an overview of how the user studies were set up and the tests conducted.

Section 6 shows the results of the user studies and discusses how well they went.

Section 7 provides a conclusion.

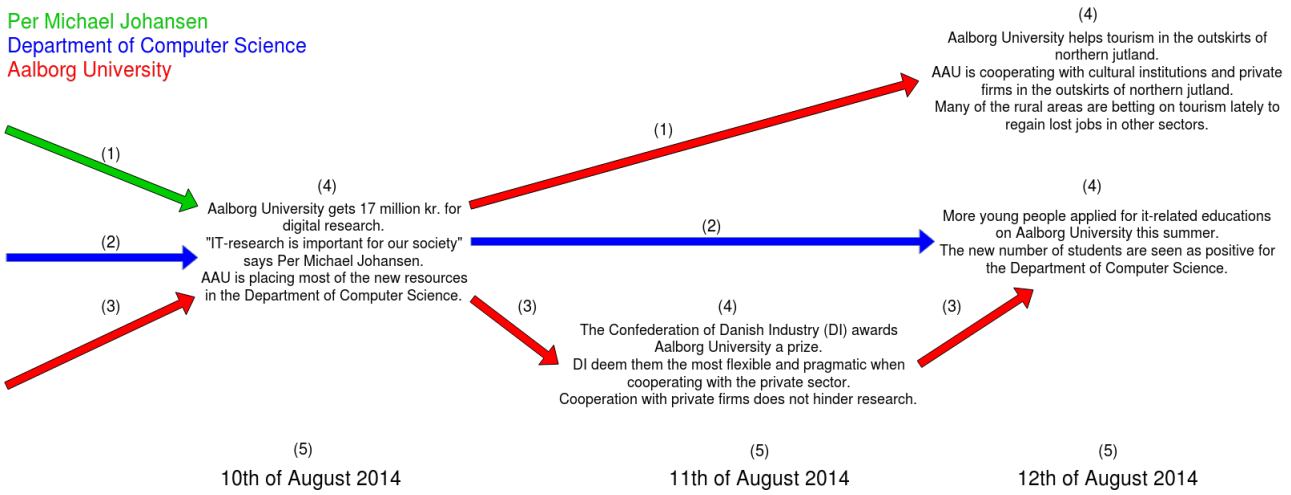

Fig. 1: Example snippet of a Metro Map focusing on Aalborg University

2. PRELIMINARIES

2.1. Terminology

We use a number of technical terms in this paper, which we define in this section.

Event: We use two different types of events: word based events $e^w = \{W, t\}$, where W is a collection of words and sentence based events $e^s = \{S, t\}$, where S is a collection of sentences. In both cases, t is a point in time which the stories described by W or S happen around, although the stories do not have to be written or mentioned at that specific date. The word based events are used in the process of creating the metro map and the sentence based events are required for visualizing the metro map.

Entity-Timeline: We define an entity-timeline as $\{seq, entity\}$ where $seq = (e_1^s, ..., e_n^s)$ and for all $e_i^s = \{S_i, t_i\}$ we have that $t_i < t_{i+1}$. That is, an entity-timeline consists of a sequence of sentence based events in chronologcal order, and an entity which all the sentence based events are related to.

Metro Map: A metro map is a set of entity-timelines and their sentence based events. This forms a map of the interactions between the entity-timelines when visualized such that timelines intersect at the same sentence based event. An example of the visualization can be seen in figure 1.

Term: In the paper, we use words, named entities as described in section 3, hypernyms and synonyms as described in section 4.4.1 to create the metro map. Terms is used as a description for words, named entities, hypernyms and synonyms.

2.2. Related Work

This section presents sources with similar ideas about timeline creation and how to represent events over time.

[2] outlines how to create metro maps over a large corpora of news. They were the main inspiration for using the metro map approach to represent timelines, but unlike our metro lines, theirs are tied to topics, which is a collection of events, e.g. a bike race taking course over several days, where each stage might be an event, rather than entities and they assume the data is from larger or several different sources. This will likely create metro maps stretching over a smaller time span. [5] provides a an overview of methods related to creating topic evolution maps, with topic evolution being how topics evolve over time, and map being the visual representation of that information. It roughly categorizes research done on topic evolution map into 2 categories: those based on probabilistic generative models and those based on non-probabilistic models. For our purpose, we have chosen a non-probabilistic model as we have been unable to find any studies comparing the efficiency of the two categories. Non-probabilistic models are models which do not rely on probability distributions and where topics often refer to a cluster of text articles or terms.

In probabilistic settings a topic is a probability distribution over terms. The models are often an extension of LDA (Latent Dirichlet Allocation) [6] [7] or PLSA (Probabilistic Latent Semantic Analysis) [8] [9]. They often incorporate the temporal aspect into the model itself. Some papers divide time into discrete time [6], which can work fine for an archive of scientific papers and news articles, while others model it as continuous time [9]. We model time as discrete as that makes sense for news articles published at discrete dates.

[10] provides a solution to detecting densely overlapping clusters. Unlike many other approaches to the issue of overlapping clusters, [10] assumes that the overlapping parts of the clusters is itself densely connected and uses this assumption in identifying the overlaps. They implement an example of a bipartite affiliation network model called BIGCLAM, which incorporates the idea of densely overlapping clusters. We use BIGCLAM to create the word based events described in section 2.1 since this clustering scheme worked well for [2] with an issue similar to ours. Since we cluster terms, the as-

sumption of densely overlapping clusters intuitively fits our problem; a word being exclusively in one cluster would result in the model only trying to fit Aalborg University into only one of the sentence based events in figure 1.

Most papers implementing a timeline or some other representation of stories over time have to consider the proper timespan of a coherent story. [1], which creates timelines for use by historians, proposes that a timespan should be as long and diverse as possible to maximize how much different information can be gained quickly by looking at the timeline. [3], which clusters news articles to reduce information overload, instead proposes a very short timespan to better represent the core of a news event. In our implementation, we do not directly need to specify a timespan for our timelines since it spans the lifetime of an entity, and as a result our model should contain long timespans such as with [1]. In section 4.5 we describe how we take time into account when creating each of the individual events. All events are described using documents originating at or near the same date.

3. NAMED ENTITIES

In the creation of a metro map, we use a Danish dataset. To find named entities in this dataset, we use the method described in [4], our preceding semester project, as well as the definition of entities described in that paper, meaning we consider any Danish proprium an entity.

As the method described in [4] has a long running time on larger datasets, we have modified the method slightly for the purpose of this paper. Those modifications are described in section 3.1.

The system described in [4] has two phases: recognition and disambiguation. In the recognition phase, all the tokens representing entities in the text are found by checking whether a similar named entity is referenced on Wikipedia and using rules common for Danish propriums, e.g. if a word is in the middle of a sentence and has a capital starting letter. In the disambiguation phase, each recognized token is paired with an entity found either on Wikipedia or LinkedIn. This is done using several scores:

- Popularity prior, which represents how much a token corresponds to an entity candidate based solely on the token's name.
- Entity-entity coherence, which represents how well an entity candidate fits in with all other entity candidates for all the recognized tokens in the text.
- Mention-entity similarity, which represents how well the context of the recognized token fits with the entity candidate.

3.1. Modifications

The original method described in [4] uses a popularity prior, mention-entity similarity and entity-entity coherence to find the disambiguation of a recognized entity.

To be able to find all entities in the entire dataset in a timeframe that is realistically useful for a metro map, we have altered the method described in [4] with a couple of rules:

Any entity that has a popularity prior with a score of 0.8 or more is disambiguated to that popularity prior. This is to prevent having to run the much more computationally heavy parts of finding mention-entity similarity and entity-entity coherence after finding the popularity prior for a large number of entities, which have some popularity prior outliers, e.g. Lego refers to the company making toy bricks in almost all cases, but a Danish arts gallery named galleri Lego exists as well. As the token Lego refers to the toy company in almost all cases, it will create a large popularity prior, but the existence of the gallery means we would calculate entity-entity coherence and mention-entity coherence as well, even though we often want the word to refer to the toy company. Although only looking at the popularity prior in this case speeds up the algorithm significantly, it also cuts off the part of the algorithm considering context and we therefore expect it to have lowered the overall accuracy. An example of how this causes a wrong disambiguation is when "Denmark" in "Denmark won the game putting them in the quarter final" is disambiguated to the country Denmark rather than the Danish sports team.

Any instance of a language is disambiguated to a country using a mapping from all languages to their native country. As tokens with countries as entity candidates have very long running times due to the large amount of references to countries on Wikipedia, we can thus limit the amount of times we encounter those. Languages are not Danish propriums and should technically not be considered entities, but since they are often recognized in the statistical named entity recognition part of the system and could be seen as a representation of their corresponding countries, we do this without loss of accuracy.

4. METRO MAP GENERATION

To create a metro map, two things are required: The metro stops, seen as (4) is figure 1 and the metro lines, seen as (1), (2) and (3) in figure 1. The process of creating those is outlined in figure 2 and described in this section.

4.1. Framework

The first step is to find a set of documents to create a metro map from. When a user wants to get an overview of any topic, we have the user do a free text search to pick documents in a news archive to base the timeline on. We call this step the *document search stage* of the framework and it is explained further in section 4.2. The documents found by this search go through a similarity comparison step, using cosine-similarity, where any documents believed not to fit into a cluster with any other documents are picked out and each get their own event. This is described in section 4.3.

We assume that we are able to identify important events in a set of documents by clustering in a bag of words scheme so that each cluster corresponds to one event. We assume

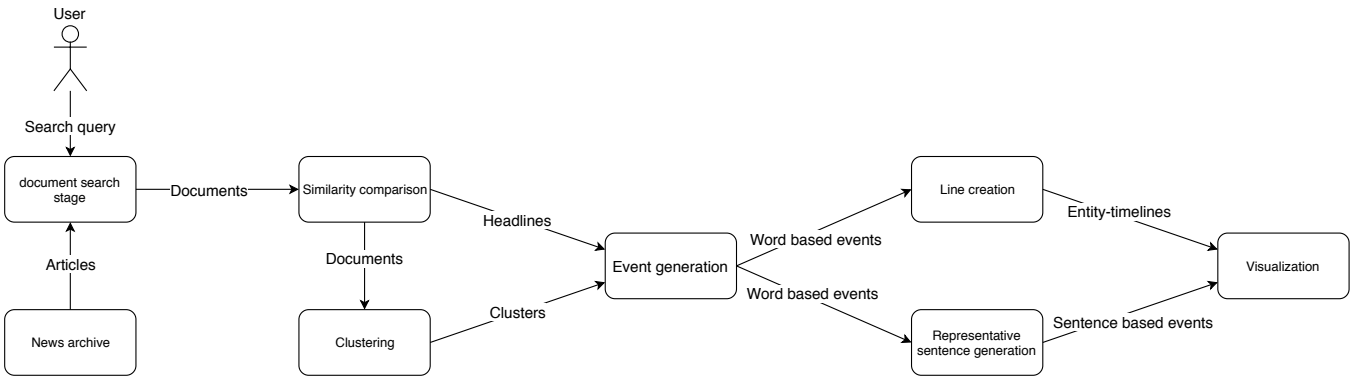

Fig. 2: System architecture

this because we note that terms in news articles describing an event often co-occur. To represent a given document we use a mix of terms, hypernyms and entities which we describe in section 4.4.

As can be seen in figure 2 we create events originating from two different stages in the framework. Some events originate from multiple documents in the *clustering*, and some originate from single documents in the *similarity comparison*. Both sets of events are used in *line creation* as described in section 4.6 and *representative sentence ranking* as described in section 4.5.

To get an overview of the progression of events a given entity is involved in, we associate a sequence of clusters to entities deemed important in the set of documents in a stage we call *line creation*. From this sequence we visually create a line between clusters. Exactly how the entities and lines are handled is described in section 4.6.

The outcome of a cluster in the *event generation* is a word based event (e^w). To represent a cluster to the user, so that the user is able to identify the event, the events are presented with sentences from the documents acquired in the *document search stage*. We select these sentences in a stage we call *representative sentence ranking*.

4.2. Document Search stage

The first stage of generating a metro map is finding the documents we need to create the stops and lines from. Any search scheme can be used for this purpose as long as it produces a set of documents. Our approach is using a user generated text query, which has to be present in each of the documents used, i.e. if the user searches for Aalborg, the word Aalborg has to be present in each of the documents used for the creation of the metro map. As the metro maps we create are very centered around entities, another option would be to focus the search on entities instead of free text. A user would be able to search through all entities present in the entire dataset and select one or more of them. One or more of the entities selected would then have to be present in any document used in the creation of the metro map, i.e. if the user selects Aalborg University, that entity would have to be present in all documents used, but the text "Aalborg University" would not have to specifically appear.

4.3. Similarity Comparison

In this stage we find events originating from a single document, in a set of documents. Say we have a set of n documents $D_{search} = d_1, ..., d_n$ found by a search query and a vocabulary V that includes terms from any given document d_i. We generate events from D_{search}, by clustering terms in V in a bag of words approach which we do based on the BIGCLAM model from [10] in section 4.4 and as is done in the metro map generation in [2]. We then represent each word-based event with the words from the generated cluster.

A problem with using this approach is that events originating in single documents are not represented in this clustering scheme. This is because terms involved in an event described only by a single document do not co-occur with terms in other documents. When dealing with a news archive with plenty of data about events like national and international oriented newspapers, this might not become a problem, because many clustering techniques would likely be able to find and rank the most important events. When dealing with a local newspaper, many events are sparsely mentioned between documents, which makes these events a concern to handle. We find these documents by making use of a cosine similarity scheme and exclude these from clustering, making them word based events originating from only 1 document.

Assume that the cosine similarity between document d_i and all other documents in D_{search} except d_i is below some threshold. We then assume that the words in d_i have so little correlation with the words in the other documents in D_{search} that we can infer d_i to be the only document describing the given event. We use tf-idf vectors as document representation for the scheme. The threshold for each of the documents in D_{search} is determined by being empirically adjusted. In our experiments we set the threshold to 0.15. We denote the set of documents we find by this method as D_{cosine}. The documents used for the clustering stage is then $D_{clustering} = D_{search} - D_{cosine}$.

4.4. Clustering

As mentioned in the last section we assume that we are able to identify important events in a set of documents by clustering in a bag of words scheme so that each cluster corresponds to one event because of co-occurrences of terms in news articles describing an event.

When clustering the documents BIGCLAM is used for event generation as in the metro map generation in [2]. We use the BIGCLAM implementation found in SNAP for python [11]. BIGCLAM models connections between terms in a graph G such that if 2 terms are mentioned in 2 or more documents, the terms are linked in $G = (E, V)$. They also introduce a background edge noise, with a probability of ϵ to add an edge between two terms which they set to $\epsilon = 10^{-8}$.

Say we have a set of documents $D_{clustering}$ and a vocabulary $V_{clustering}$ that includes terms from any given document d_i in $D_{clustering}$. We want to find the clusters $C = c_1, ..., c_K$, with K being the number of clusters and $c_i \subseteq V_{clustering}$. In BIGCLAM the number of clusters K is a fixed value, and we describe later in this section how we determine that. BIGCLAM assumes the probability of terms u and v being linked as

$$p(u, v) = 1 - exp(-F_{u,:} \cdot F_{v,:}^T) \tag{1}$$

where F is a latent $|V_{clustering}| \times K$ matrix with F_{wc} being the strength of connection between term w and a latent cluster c. The aim of BIGCLAM is to estimate F and use strength of connection between a term and a latent cluster F_{wc} to evaluate whether the term belongs to cluster c. The probability of generating a graph G is then

$$P(G|F) = \prod_{(u,v)\in E} p(u, v) \prod_{(u,v)\notin E} (1 - p(u, v))$$

We estimate the latent matrix F by maximizing the log likelihood: $\arg\max_{0\leq F} ln(P(G|F))$ in a gradient ascent scheme and pick the highest scored terms for each cluster above some threshold δ. Let $l(F) = ln(P(G|F)$. Then the update function in the gradient ascent scheme can be determined by

$$\nabla l(F_u) = \sum_{v\in N(u)} F_v \frac{exp(-F_{u,:} \cdot F_{v,:}^T)}{1 - exp(-F_{u,:} \cdot F_{v,:}^T)} - \sum_{v\notin N(u)} F_v,$$

where N(u) is the set of neighbors of node u. BICLAM finds that setting $\delta = \sqrt{log(1 - \epsilon)}$, with ϵ being the background noise probability, works well in practice. We determine the number of clusters K by running BIGCLAM sessions with different K-values and we choose the K that results in the lowest log likelihood. In our case we tested on all K-values between 1 and half the number of documents. This is because any cluster should contain words from at least two documents.

4.4.1. Quality Improvements

To improve the quality of the clusters done by BIGCLAM we process the the input documents $D_{clustering}$ and their corresponding vocabulary $V_{clustering}$. We do not use all words in each document in $V_{clustering}$, because not all words are equally important for the document. Instead we use the top 50 scored words, with tf-idf as a metric, as well as a stopword list [12] that excludes the 350 most used words in Danish. When encountering named entities in a document, we append the stop words to the document representation along with the top 50 scored words, and if we have been able to disambiguate the entities as described in section 3, we append the disambiguated entity otherwise we append the recognized one. This is because we want an entity to be treated equally regardless of the different ways to address the same entity. We then lemmatize the words using CST's lemmatizer [13] so that different inflections of a word across documents are also treated as equal. We also introduce hypernyms and synsets, obtained from DanNet [14], such that all hypernyms or synonyms of the top 50 scored words are included in the document representation. A hypernym is a word generalizing a set of words. The words in this set are called the hyponyms of the generalizing word. As an example we could say the verb 'look' is a hypernym for the verbs 'stare' and 'view'. In this example 'stare' and 'view' are then hyponyms of 'look'. We introduce these hypernyms and synonyms such that the functional uses of the words can be treated equally between documents where it normally would not.

4.5. Representative Sentence Ranking

To better show the events that each cluster represents, we represent its contents through sentences from the documents in the news archive. This is shown in the example in figure 3. To do this, we have created a sentence ranking system based on ideas from [15], although we have simplified the process to fit our problem.

Two kinds of events need to be represented, the ones originating from only one document, which are found in similarity comparison as seen in figure 2 and the ones originating from multiple documents, which are found in clustering as seen in figure 2.

For the events originating from a single document, the article's headline and the first 2 sentences are used as the sentence representation. This is solely because our other events have a 3-sentence representation and we do not want to confuse the user with an arbitrary amount of sentences per event. If the document has less than 3 sentences the maximum amount of sentences in the document is used.

For the events originating from multiple documents, the representative sentence ranking system requires an input of words, hypernyms/synonyms and named entities, named the token input. These represent which parts of the documents we find important. We construct this input based on the clusters created during event generation. Each cluster is a set of the most co-occurring terms from document representa-

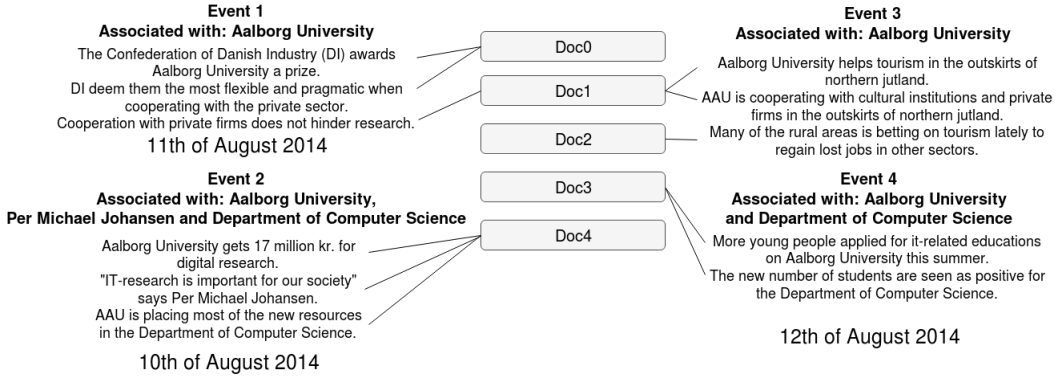

Fig. 3: We represent clusters as events through sentences picked from documents in the news archive

tions in the news archive as discussed in section 4.4. The terms for the summarization system for each cluster are created from those words, hypernyms/synonyms and named entities. An example of this could be the token input for the first event represented in figure 1: {e.Aalborg University, e.Per Michael Johansen, e.Department of Computer Science, research, h.Material}. Here e denotes entities and h denotes hypernyms/synonyms.

In addition to the token input from the cluster, the system also needs documents from which it can extract a summarization based on the token input. For this purpose, all documents found in the document search stage are used.

Each sentence in each document starts with a score of 0 and is then compared to the input from the cluster. If the sentence contains a word, the hyponym of a hypernym or a synonym found in the input, 1 is added to its score. For terms in the input which are entities we define a function $DF : e \rightarrow \mathbb{Z}$ which maps an entity to the number of documents that contain it, based on documents found in the document search stage. If a sentence contains an entity found in the input, a weight multiplied with the inverse of the entity's document frequency score ($DF(e)^{-1}$) is added to the sentence's score. The reasoning behind this is that the words used have already been selected through the tf-idf scheme in the clustering step, suggesting all words are relevant to the cluster. We do this because we assume that entities that seldom appears in all documents are the most important actors in the event. We also want the entities most relevant to the cluster to control what sentences are selected. Thus the entity weight must be relatively high, in our case we used 20, as it both has to make up for the reduction of score by the inverse of the document frequency compared to words and hyponyms/synonyms and signify the importance of entities.

Using the input from the cluster, the words W, the entities E and the hyponyms H of a hypernym are searched for in each sentence $sent$ in the documents. The score for each sentence is then calculated as in equation 2 where $n_{w/h/e}$ is the text representing each part of the input, s a sentence and ew the entity weight.

$$score(s) = \sum_{w \in W} \mathbf{1}_s(n_w) + \sum_{h \in H} \mathbf{1}_s(n_h) + \sum_{e \in E} \mathbf{1}_s(n_e) * ew * DF(e)^{-1} \quad (2)$$

Once the sentence scores have been found, time is factored in. This is to reflect the nature of an event happening at a specific time point. If we have two similar events happening a year apart, we still wish to present those as separate events rather than one simply because the terms contained in their clusters are alike. We start by finding a primary time point, which is when we estimate the event taking place. To do this we define an article score for a specific event to be the sum of all scores for the sentences contained in that article. The primary time point is then found by scoring the dates of each article written on that date by that article's score, e.g. if we have 3 articles, one with a score of 7 on 10-10-2010, one with a score of 1.3 on 10-10-2010 and one with a score of 8 on 12-12-2010, 12-12-2010 gets a score of 8 and 10-10-2010 gets a score of 8.3 and is elected the primary time point. Once a primary time point is found, each sentence not on that time point has its score altered depending on how far away it is from the primary time point. This is done as seen in equation 3, where s is a sentence and pt is the primary time and t is a time-factor with $0 < t < 1$.

$$timedScore(s) = score(s) * t^{|date(s) - pt|} \quad (3)$$

Here we define a date minus another date to give the result as the difference of amount of days between them. The highest scoring sentences, with time taken into account, are then selected to represent the event. The amount of sentences selected can be adjusted depending on the information need of the user.

4.6. Entity-Timeline Generation

All unique entities found during *Event generation* are associated with a timeline representing a sequence of coherent events that entity was involved in. This is the task of the *Line creation* module and how this is done will be described in this section.

Sentence	Sentence score
AAU is placing most of the new resources in the Department of Computer Science.	5.50
Aalborg University gets 17 million kr. for digital research.	3.00
The Confederation of Danish Industry (DI) awards Aalborg University a prize.	2.00
Aalborg University helps tourism in the outskirts of northern jutland.	1.00
AAU is cooperating with cultural institutions and private firms in the outskirts of northern jutland.	0.50
$event\_score(AalborgUniversity)$	12.00

Table 1: This is an example calculating $event\_score(AalborgUniversity)$ for the example snippet in figure 1 (Alborg University is present in all the example sentences shown here)

To construct a timeline we need a sequence of events in ascending order based on date. To do this we need to associate each event with a specific date. From the *representative sentence generation* module we get a series of word based events associated with sentences. Each sentence is associated with a specific article from the *News archive* as seen in figure 3. In *Line creation* we associate articles with sentence based events containing their corresponding sentences. We find the date for an event based on the dates when the associated articles were written. You could associate the event with the earliest date or you could choose the article associated with the best scored sentence from the *representative sentence generation* module. We chose the latter approach because we found that just choosing the earliest date tends to result in dates for sentences describing a future event.

In the beginning of *Line creation* the amount of entities associated with a timeline is the amount of unique entities mentioned in all the word based events. *Line creation* then limits these lines per event by removing entities from the event based on a series of conditions. Since lines are associated with an entity, reducing entities in an event limits the amount of lines that passes through that event. These conditions are listed below:

1. We set each event to be associated with a maximum amount of entities. These entities are then presented to the user as being part of the event. This ensures that we get the most important entities involved in the event, and to reduce the amount of information presented to the user.

   We find this by defining the function $event\_score(entity)$ which associates each entity with a score. $event\_score(entity)$ is the sum of the scores of all sentences, across all events, which $entity$ is a part of. The sentences are scored in the *representative sentence generation* module. With this score we choose for each event only the highest scored entities. An example of this can be seen in table 1.

   We do this to keep the entities mentioned in the sentences, which makes more sense for the user since that is what the user is going to see. The maximum amount of entities per event is set empirically based on personal evaluations of the information load presented to the user. In our experiments we have set it to be 4.

2. We discard entities that are only associated with one specific date, since those entities do not tell a story progression involving that entity

3. We set each specific date to be associated with a maximum amount of events. This is to reduce the complex structure of many interacting entities. Limiting the amount of events per date also limits the amount of entity-timelines passing each date. This maximum is also set empirically based on personal evaluations of the information load presented to the user. To ensure that we keep only the most important enitity-timelines we rank events of the metro map for each date. This is done as shown in figure 4, by recognizing and disambiguating entities of documents associated with the entire metro map, and sorting them according to the number of times they are mentioned. When the number of events of a given date crosses the threshold, the events which contains the most mentioned entities are kept. In our experiments we have set the maximum amount of events per date to be 3.

4.6.1. Splitting Entity-Timelines

When events with the same entity occur at the same date, we split the line, such that it represents temporally parallel events that entity is associated with. An example of this is seen in figure **??**, where Aalborg University is involved in two separate events on the 12th of August 2014. This introduces a new problem, because it is undefined which of the split entity-timelines future events shall be connected with. To get a coherent event progression we associate each event with the line which is most similar in cosine-similarity based on the words associated with that event. To minimize the number of entity-timelines per entity, we allow lines which have stopped progressing to continue with a new story progression involving the same entity.

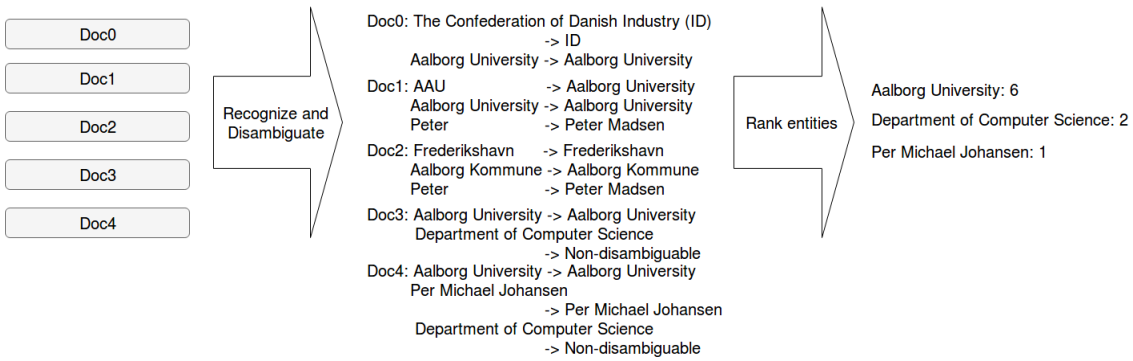

Fig. 4: Example of ranking entities by disambiguating documents from example of fig. 1 and fig. 3.

5. EXPERIMENTS

5.1. Experimental setup

Overall	
Number of documents	98088
Number of recognitions	760001
Number of disambiguations	508458
Avg. size of document (in words)	150.3
Avg. recognitions per document	7.7
Avg. disambigations per document	5.1

Table 2: Statistics for the NORDJYSKE news archive from 07/06/2017 to 02/27/2018

For measuring how well our system performs, we conduct user evaluated tests on 9 testers to check how coherent the different parts of the metro map are and a test to see how coherent the lines intersecting the events are with the events.

The dataset used for testing is a news dataset provided by NORDJYSKE, which contains 98088 news articles in Danish focused on local stories in the area around Aalborg. Some of the articles are subsets of each other, but none are duplicates. They are extracted in the timespan from 07/06/2017 to 02/27/2018. Statistics about the dataset used from NORDJYSKE can be seen in table 2. Statistics about documents extracted from each search string can be found in appendix B.

To search through the dataset (*the document search stage* in figure 2), we use a string as a query and retrieve all documents containing that string. The strings used are *"Socialdemokratiet"*, *"Birgit Hansen"*, *"AAB"*, *"Industri"*, *"Aalborg Portland"*, *"Frederikshavn"*, *"Karneval"*, *"Thomas Kastrup"*, *"Aalborg Kommune"* and *"Per Michael Johansen"*.

Each of the components described in section 4 have been programmed using Python 3.6. For implementing BIGCLAM, which is used for clustering, we use the version found in SNAP for python [11].

The test instructions presented to the user can be seen in appendix C. The evaluation tables provided for the test to contain the user's answers have been omitted from the appendix.

In the rest of this section, we will present the metrics used for measuring how effectively and coherently the metro map conveys information. We have been unable to find standard measurements for this kind of work, but some of our metrics are similar to that of other papers creating timelines. We do not compare ourselves to those papers, as our dataset and focus on entities are quite different from theirs.

5.2. Event coherence

For testing the event coherence, the user is shown 5 sentence based events per search string, except for the search string 'Frederikshavn' which only has 4. The users are then asked to rate how well each sentence from the sentence based event fits into the event on a scale of 1-5; 1 suggesting it does not fit in at all, 5 suggesting it fits perfectly. If the user is unable to figure out what the event is about at all, and hence cannot decide whether any sentences fit in, all sentences in that event are given a score of 1. There is a total of 130 sentences for the user to rate, which were chosen randomly from clusters in each search string. These sentences are shown in appendix D.

Once scores have been obtained, we calculate the average score for each event, which we call the event coherence (EC) score. We also calculate Fleiss' kappa, denoted κ, to get a perspective on the user agreement.

5.3. Entity-event coherence

The coherence of the entities and events, which we call the entity-event coherence, is tested by presenting the user with an event and the entities of the entity-timelines running through it, which they rate with one of 3 options:

Correct, when they are sure the entity, that the entity-timeline represents, is involved with the event i.e. for the first event in figure 1, this could be Aalborg University.

Possible, when they cannot tell for sure if the entity is involved in the event, but it might be there i.e. for the first event in figure 1, this could be Cassiopeia (The Department of Computer Science's building).

Incorrect, when they do not think the entity is involved in the event, i.e. for the first event in figure 1, this could be Louis XIV of France.

If the event itself is too incoherent to understand, all the entities representing the entity-timelines intersecting the event are considered incorrect.

We use two metrics: Strict entity-event coherence (SC) and lax entity-event coherence (LC) using the amount of entities categorized as correct (c), the amount of entities categorized and possible (p) and the amount of entities categorized as incorrect (i). We calculate $SC = \frac{c}{c+i}$ and $LC = \frac{c+p}{c+p+i}$. An entity-event coherence close to 1 suggests that the entities are each connected to events relevant to them, whereas an entity-event coherence close to 0 suggests very few entities connected to the events are relevant to them. The reasoning behind using two measures is that we can incorporate the benefit of the doubt in LC, where possible entities are considered correct and have the more constricting SC, where possible entities are considered incorrect. In addition to those measures, we also calculate Fleiss' kappa to find the agreement between the users.

5.4. Entity-Timeline familiarity difference

In addition to the measurements of individual events, we wish to gain an overview of how much the users have learned after reading the metro map. We have implemented a user interface in Java to navigate the metro map for this purpose. This implementation is detailed in appendix A. We ask them how well they feel they know the subject matter, that is the search string, on a scale of 1-5 before showing them the metro map and again after showing the metro map. 1 represents no knowledge of the subject in the timespan and 5 represents complete knowledge of the subject in the timespan. We then use the average increase/decrease in rating as the entity-timeline familiarity difference (Δf) score, to estimate how much the user learned about the subject from reading the metro map.

6. RESULTS

In this section, we present and describe our findings from the user evaluation tests described in section 5. Table 3 gives an overview of the abbreviations used in this section.

EC	event coherence
κ	Fleiss' kappa
SC	strict entity-event coherence
LC	lax entity-event coherence
c	correct entities
p	possibly correct entities
i	incorrect entities
Δf	entity-timeline familarity difference
fb	familiarity before
fa	familiarity after

Table 3: Abbreviations

	κ
Event coherence	0.09
Entity-event coherence	0.49

Table 4: Fleiss' kappa

6.1. Event coherence

	Score
EC	3.44

Table 5: Event coherence

Score	Amount
1	102
2	122
3	331
4	393
5	222

Table 6: Event coherence scores

From table 5 we see that we get an event coherence of 3.44, which considering it is a score between 1 and 5 where 5 is perfect suggests many events are at least coherent enough to be understood, but some sentences do not fit in particularly well. It should be noted though, as can be seen in table 6, that the majority (81%) of the rating were a score of 3 or more, suggesting the sentence might not be the best fit, but it is not entirely wrong neither.

In table 4 we see the κ for event coherence of 0.09, which is quite low. This might be because of the way the test is constructed, where the user first has to decide what the event being represented is about and then how well each sentence fits into that event, which can create several different interpretations of what sentences are right and wrong. Another reason for the low κ might be that it does not consider how closely related two categories are, e.g. a user answering 1 and a user answering 2 are in closer agreement than a user answering 1 and a user answering 5, which is a nuance that Fleiss' kappa does not take into account. We can illustrate this by grouping some of the close categories such that scores from category 1 and 2 are merged and scores from category 4 and 5 are merged. This results in a κ of 0.21.

6.2. Entity-event coherence

c	p	i	SC	LC
559	74	249	0.69	0.72

Table 7: Entity-event coherence

As seen in table 7 we achieve a strict entity-event coherence of 0.69 and a lax entity-event coherence of 0.72. Both scores

reflect that the users consider a lot more of our entities to be correctly placed on the events than incorrectly. About 70% of the scores are considered to be correctly placed according to both scores.

The κ is 0.49 for entity-event coherence, as seen in table 4. According to [16] this suggests a moderate agreement between raters. Here we do not see the same problem as with event coherence because the ratings are not depending on continuous numbers in the same way, although possibly correct entities are more related to both correct entities and incorrect entities than correct and incorrect entities are to each other.

6.3. Entity-timeline familiarity difference

Entity Name	fb (avg.)	fa (avg.)	Δf
Aalborg Portland	1.00	2.33	1.33
AAB	1.56	2.78	1.22
Birgit Hansen	1.00	2.78	1.78
Frederikshavn	1.56	2.22	0.66
Industri	1.44	2.44	1
Karneval	2.11	2.56	0.45
Per Michael Johansen	1.00	2.44	1.44
Socialdemokratiet	2.11	2.89	0.78
Thomas Kastrup	1.11	2.78	1.67
Average	1.43	2.58	1.15

Table 8: Entity-timeline familiarity difference

The entity-timeline familiarity difference can be seen in table 8, where we have an average Δf of 1.15. As complete familiarity with a subject is 5, this means that they rate themselves to be familiarized with $\frac{1.15}{5} = 23\%$ more of the subject than before.

If we assume the knowledge contained in the news archive is rated 5 when learned, which is possible for some search strings, we can say that the gap between average fb (familiarity before) of 1.43 and 5 indicates the amount of information that can possibly be learned. Given this assumption we optimally want the users to be familiarized with $\frac{5-1.43}{5} = 71\%$ more of the subject than before, rather than 23%. This shows that in the effort to reduce information overload, relevant information is likely lost.

The average fa (familiarity after) of 2.58 indicates the users on average have mediocre knowledge about the subject after going through the metro map. Considering all entities have a fa of at least 2 and all of them have at least some Δf (familiarity difference), we see that the metro map conveys some information to the user for all subjects.

In practice, we might expect more users to have more average fb about the subjects they search for, as the current average of 1.43, which is almost no knowledge about the subject, would make it unlikely that they even knew to search for the subjects to learn more. This will likely also change Δf. Either it will be higher as they understand more events in the metro map, or it will be lower because the metro map now only conveys information that they already know.

The individual results for Δf in table 8 do suggest that the metro map is better suited for some search queries than others. Especially names of people in the community, in our case *Birgit Hansen* with a Δf of 1.78, *Per Michael Johansen* with a Δf of 1.44 and *Thomas Kastrup* with a Δf of 1.67 yield a higher Δf compared to the 1.15 average. This could be due to a tendency of people often being mentioned in articles in relation to their profession, e.g. Per Michael Johansen (the rector of Aalborg University) is often mentioned and discussed in relation to the university, whereas something like Aalborg Portland (a company originating from Aalborg) more often than not results in articles about Aalborg Portland Park, a football stadium that they sponsored.

The search queries that result in many documents, e.g. *Frederikshavn* with 4101 documents and *Socialdemokratiet* with 1087 documents, as seen in table 16 and 17 in appendix B, seem to have a tendency of providing worse results. Both end up with a Δf of less than 1. A search query such as *AAB* with 721 documents as seen in table 15 in appendix B results in a better Δf of 1.22 though, so more documents is not necessarily a problem, but it could seem that there is a point where you get too many and the results suffer from it. We have been unable to find out whether this is the case for other search string resulting in many documents.

The very broad search queries also seem to do worse than searching for specific entities, e.g. *Karneval* (Danish for carnival) has a Δf of 0.45 and *Industri* (Danish for industry) has a Δf of 1, which is better, but still lower than other search strings used. This could likely be a reflection of the metro map focusing on entities and any search string that is not an entity cannot have its own entity-timeline and can therefore be harder to follow.

7. CONCLUSION

We introduced a system that reduces a news archive to a series of events linked in timelines representing entities. This was done to reduce information overload. We found that the system was able to help people learn about specific local topics, but that the system did not entirely remove information overload. We noticed that a lot of the users found there was superfluous information in most of the metro maps. Cutting this information away is intended to be handled by the system, which suggests that some components still need to be tuned or more components need to be added. The users considered many of the events and entity-timelines to be understandable and coherent, so a new metro map with more restrictions on what events to include and how to link them together might improve a lot on the initial issue.

Another component which is obvious to improve is using a more sophisticated document search stage, which would likely greatly improve the quality of the metro map presented to the user or switching the representative sentence generation component to a more sophisticated method e.g. looking at one of the graph-based approaches in topic summarization.

8. REFERENCES

[1] J. Singh, W. Nejdl, and A. Anand, "History by Diversity: Helping Historians Search News Archives," in *CHIIR '16*, pp. 183–192, 2016.

[2] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, and J. Leskovec, "Information Cartography: Creating Zoomable, Large-scale Maps of Information," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), pp. 1097–1105, ACM, 2013.

[3] S. Vadrevu, C. H. Teo, S. Rajan, K. Punera, B. Dom, A. J. Smola, Y. Chang, and Z. Zheng, "Scalable Clustering of News Search Results," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, (New York, NY, USA), pp. 675–684, ACM, 2011.

[4] M. Stisen and M. H. Thygesen, "Named Entity Recognition and Disambiguation for the Danish Language," 2018.

[5] H. Zhou, H. Yu, R. Hu, and J. Hu, "A survey on trends of cross-media topic evolution map," *Knowledge-Based Systems*, vol. 124, pp. 164 – 175, 2017.

[6] A. Ahmed and E. Xing, *Dynamic Non-Parametric Mixture Models and The Recurrent Chinese Restaurant Process: with Applications to Evolutionary Clustering*, pp. 219–230.

[7] D. Kim and A. Oh, "Topic chains for understanding a news corpus," in *Computational Linguistics and Intelligent Text Processing* (A. Gelbukh, ed.), (Berlin, Heidelberg), pp. 163–176, Springer Berlin Heidelberg, 2011.

[8] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, (New York, NY, USA), pp. 198–207, ACM, 2005.

[9] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in *2011 IEEE 11th International Conference on Data Mining*, pp. 378–387, Dec 2011.

[10] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 587–596, ACM, 2013.

[11] J. Leskovec, "SNAP." `https://snap.stanford.edu/snap/description.html`.

[12] hermitdave, "FrequencyWords." `https://github.com/hermitdave/FrequencyWords`.

[13] B. Jongejan, "CST." `https://github.com/kuhumcst/cstlemma`.

[14] Det Danske Sprog og Litteraturselskab, "DanNet." http://www.wordnet.dk/menu%3Fitem=2.html.

[15] J. Steinberger, K. Jezek, and M. Sloup, "Web Topic Summarization," in *Conference on Electronic Publishing*, pp. 322–334, ELPUB, 2008.

[16] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

Appendices

Fig. 5: Screen shot of the user interface

A. USER INTERFACE FOR USER EVALUATION

Here will be a description of the GUI implementation presented to the user as part of the user evaluated tests described in section 5. An example of the user interface can be seen in figure 5. It is constructed using Java 2D Graphics. A search bar exists in the upper left corner of the user interface, which can be used to input the search query described in section 4.2. Once the user has searched, a number of circles appear. Those circles are the events described in section 4.5. Each circle is linked to the others with a number of lines colored to represent the entity-timelines described in section 4.6. The entities assigned to the events appear just above the events in text with a color matching its corresponding entity-timeline. In addition, a legend of all entities appearing in the metro map can be found in the upper left corner of the user interface with its corresponding color.
If the user hovers the mouse over any of the presented events, the sentence representation of it will appear in a text box as seen in figure 5.
Only the presentation of the metro map to the user is created using Java, the rest of the system, i.e. all parts described in section 4, are created using Python. The end result of the Python program is a text document used by the Java program to generate the visual metro map.

B. STATISTICS

This section contains tables of statistics of each of the search queries used in the experiments. What are listed is the number of documents the search query resulted in, the number of named entities recognized in them, the number of named entities disambiguated in them and the number of unique entities recognized and disambiguated in them as well as statistics about average amount of words per document, average named entity recognitions per document and average named entity disambiguations per document.

Aalborg Portland	
Number of documents	67
Number of recognitions	1173
Number of disambiguations	910
Number of unique recognitions	733
Number of unique disambiguations	438
Avg. size of document (in words)	346.5
Avg. recognitions per document	17.5
Avg. disambigations per document	13.6

Table 9: Statistics for the search string: Aalborg Portland

Birgit Hansen	
Number of documents	172
Number of recognitions	2861
Number of disambiguations	2012
Number of unique recognitions	1375
Number of unique disambiguations	656
Avg. size of document (in words)	318.2
Avg. recognitions per document	16.6
Avg. disambigations per document	11.7

Table 10: Statistics for the search string: Birgit Hansen

Industri	
Number of documents	360
Number of recognitions	5356
Number of disambiguations	3757
Number of unique recognitions	3015
Number of unique disambiguations	1601
Avg. size of document (in words)	411.8
Avg. recognitions per document	14.9
Avg. disambigations per document	10.4

Table 11: Statistics for the search string: Industri

Per Michael Johansen	
Number of documents	19
Number of recognitions	315
Number of disambiguations	240
Number of unique recognitions	167
Number of unique disambiguations	104
Avg. size of document (in words)	470.9
Avg. recognitions per document	16.6
Avg. disambigations per document	12.6

Table 12: Statistics for the search string: Per Michael Johansen

Thomas Kastrup	
Number of documents	410
Number of recognitions	6135
Number of disambiguations	4466
Number of unique recognitions	2019
Number of unique disambiguations	980
Avg. size of document (in words)	369.6
Avg. recognitions per document	15.0
Avg. disambigations per document	10.9

Table 13: Statistics for the search string: Thomas Kastrup

Karneval	
Number of documents	46
Number of recognitions	1084
Number of disambiguations	697
Number of unique recognitions	543
Number of unique disambiguations	302
Avg. size of document (in words)	405.1
Avg. recognitions per document	23.6
Avg. disambigations per document	15.2

Table 14: Statistics for the search string: Karneval

AAB	
Number of documents	721
Number of recognitions	9603
Number of disambiguations	7184
Number of unique recognitions	3672
Number of unique disambiguations	1709
Avg. size of document (in words)	240.8
Avg. recognitions per document	13.3
Avg. disambigations per document	10.0

Table 15: Statistics for the search string: AAB

Frederikshavn	
Number of documents	4101
Number of recognitions	53608
Number of disambiguations	33984
Number of unique recognitions	16095
Number of unique disambiguations	6267
Avg. size of document (in words)	182.2
Avg. recognitions per document	13.1
Avg. disambigations per document	8.3

Table 16: Statistics for the search string: Frederikshavn

Socialdemokratiet	
Number of documents	1087
Number of recognitions	18443
Number of disambiguations	14155
Number of unique recognitions	5809
Number of unique disambiguations	2786
Avg. size of document (in words)	404.7
Avg. recognitions per document	17.0
Avg. disambigations per document	13.0

Table 17: Statistics for the search string: Socialdemokratiet

C. EVENT COHERENCE EXPERIMENT

This section contains the text we presented the user as part of the user evaluated tests. We have omitted the tables where the users wrote their evaluations. Otherwise it contains a description of their task as well as the events they were asked to evaluate:

Firstly we would like you to rate how well you know each of the following subjects on a score from 1-5. Consider each subject only in the local context of Aalborg, Frederikshavn and nearby cities and only within the timespan 6th of July 2017 - 27th of February 2018. Once each of the subjects have a score, we will present a program for you, which you get to explore, which should hopefully provide aditional information about the subjects. Once you have explored the program for each of the subjects, we ask you to write down a new score from 1-5 representing how much you now know about the subject. The subjects are:

- Aalborg Portland
- AAB
- Birgit Hansen
- Frederikshavn
- Industri
- Karneval
- Per Michael Johansen
- Socialdemokratiet
- Thomas Kastrup

Once you have scored each of the 9 sentences above, you will be presented with a number of text boxes describing an event, which is done using 3 sentences, those can be seen in appendix D. If you need more information on the meaning of any of the words or sentences presented to better understand the event, we will answer to the best of our ability or provide you with documents detailing their meaning.

For each event you are presented with, you should rate how well each of the 3 sentences fit in with the others and rate them accordingly on a scale of 1-5, 5 being fits well and 1 being does not fit at all. If none of the sentences fit together in any way, tell us and we will move on to the next event.

An example of a good event, where the clusters fit well together, could be "Earning Denmark yet another gold medal in the Olympics", "Her performance was exemplary" and "The result was unexpected".

An example of a bad event, where none of the sentences fit together could be "The president won a landslide victory", "Earning the country yet another gold medal in the Olympics" and "The police were on the scene within minutes".

In addition to rating the sentences, each event has a number of entities attached to it, which are written above the sentences. We ask you to rate each of those entities as either:

- Correct, when the entity fits well with the sentences.
- Possible, when you think the entity might fit with the sentences.
- Incorrect, when the entity has nothing to do with the sentences.

To illustrate we give the following example: Given an event with the 3 sentences "Earning Denmark yet another gold medal in the Olympics", "Her performance was exemplary" and "The result was unexpected".

One example of a correct entity for the event could be "Denmark".

An example of a possible entity for the event could be "Per Olesen" (an Olympic judge, ask us if you are in doubt about who or what any of the mentioned entities are).

An example of an incorrect entity could be "England", which may have something to do with the Olympics, but has nothing to do with a Dane winning a medal in it.

D. EVENTS USED FOR TESTING

Portland 43 (Aalborg, Viborg).png Portland 43 (Aalborg, Viborg).bb

clusternr: 43*
1: INDLAND: Dansk Fjernvarme tror ikke, at en afgørelse fra Skatteankenævnet om afgiftsfrihed for overskudsvarme fra Facebooks nye server-center i Odense får betydning for eksisterende aftaler om overskudsvarme fra virksomheder til fjernvarme.
2: Portland-direktør Michael Lundgaard Thomsen har ikke fordybet sig i afgørelsen fra Skatteankenævnet, men han har siden sin første melding erkendt, at den næppe får betydning i Aalborg..
3: Aalborg Portlands direktør, Michael Lundgaard Thomsen, udtrykte ellers umiddelbart efter afgørelsen optimisme med hensyn til at udnytte en endnu større del af virksomhedens overskudsvarme i det aalborgensiske fjernvarmesystem, og miljø- og energirådmand Lasse P.

Portland 44 (Aalborg).png Portland 44 (Aalborg).bb

clusternr: 44*
1: INDLAND: Dansk Fjernvarme tror ikke, at en afgørelse fra Skatteankenævnet om afgiftsfrihed for overskudsvarme fra Facebooks nye server-center i Odense får betydning for eksisterende aftaler om overskudsvarme fra virksomheder til fjernvarme.
2: Olsen (EL), efter at Skatteankenævnet har givet Facebook medhold i, at der ikke skal betales afgift på den overskudsvarme, selskabets datacenter skal levere til Odense.
3: - I de igangværende forhandlinger, vi har med Aalborg Portland, har afgiften været et stort problem, så med Skatteankenævnets afgørelse er der ryddet en stor sten af vejen hen mod en aftale, siger rådmanden, der uddybende forklarer, at en varslet afgift fra statens side ville øge varmeregning for et standardhus med 350 kroner årligt med den nuværende mængde overskudsvarme i fjernvarmenettet.

Portland 45 (Rasmus Wurtz, Superligaen, Aab Fodbold, Aalborg).png Portland 45 (Rasmus Wurtz, Superligaen, Aab Fodbold, Aalborg).bb

clusternr: 45*

1: AALBORG: Skaden i baglåret er han kommet sig over - alligevel er AaB-anføreren Rasmus Würtz ikke med i truppen til dagens kamp i ALKA Superligaen, hvor AaB tager imod FC Helsingør på Aalborg Portland Park.

2: - Jeg vil da umiddelbart sige, at der var mange andre gange i de sidste seks uger, hvor en omgang influenza havde passet bedre ind i programmet, lyder det med et lille smil fra Rasmus Würtz, der ikke har været med for AaB siden han fik en forstrækning i baglåret i kampen mod AC Horsens 18.

3: Rasmus Würtz er så småt tilbage, men han måtte udeblive fra kampen på grund af sygdom.

Portland 50 (Aalborg, Robert Kakeeto, FC Helsingor, Aab Fodbold, Filip Lesniak).png Portland 50 (Aalborg, Robert Kakeeto, FC Helsingor, Aab Fodbold, Filip Lesniak).bb

clusternr: 50*

1: Så selvom Filip Lesniak og Robert Kakeeto til tider har imponeret, så virker det ikke som en fuldgod erstatning for den foretrukne midtbaneduo, der tilsammen har mere end 700 AaB-kampe at trække på.

2: Dermed var det Robert Kakeeto og Filip Lesniak, som bemandede maskinrummet for AaB.

3: Filip Lesniak har været fast starter længe, og også Robert Kakeeto fra Uganda har leveret gode præstationer på pladsen.

Portland 55 (Aalborg, Jakob Ahlmann, Aab Fodbold, Yann Rolin, Pavol Safranko).png Portland 55 (Aalborg, Jakob Ahlmann, Aab Fodbold, Yann Rolin, Pavol Safranko).bb

clusternr: 55*

1: Spørgsmålet er så, om ammunitionen er klar i samme omgang, for hverken Jannik Pohl eller Pavol Safranko overbeviste om, at de er klar til for alvor at bevæge sig mod toppen af Superligaens topscorerliste.

2: Jannik Pohl har kæmpet med et ømt knæ, der gør, at han mangler de sidste fem procent af hans helt store force, farten, og Pavol Safranko er stadig bedre med ryggen til mål, end han er inde i boksen.

3: Jakob Ahlmann presser sig på til positionen som venstre back, Pavol Safranko er en mulighed på bekostning af Jannik Pohl i angrebet, Yann Rolim konkurrerer med Frederik Børsting om venstrekanten, og endelig er det meget tæt mellem Patrick Kristensen og Kristoffer Pallesen på positionen som højre back..

8 (Aab Fodbold, Aalborg).png 8 (Aab Fodbold, Aalborg).bb

clusternr: 8

1: aab nedjusterer i overskuddet.

2: aalborg: fodboldklubben aab nedjusterer forventningerne til årets resultat fra et overskud før skat i niveauet 25 millioner kroner til et overskud før skat i niveauet 15-20 millioner kroner.

3: det skriver selskabet aalborg boldspilklub a/s i en fondsbørsmeddelelse.

66 (Aalborg).png 66 (Aalborg).bb

clusternr: 66

1: 70 år.

2: ▫ søren rasmussen, cedervænget 5, aalborg, fylder mandag 70 år.

3: søren rasmussen er født i aalborg og har boet i byen hele sit liv.

78 (Aab Fodbold).png 78 (Aab Fodbold).bb

clusternr: 78

1: aab-talent fik landsholdsdebut på u16.

2: ringkøbing: tirsdag spillede det danske u16-landshold en venskabskamp mod schweiz' ditto.

3: her debuterede aab's casper gedsted.

100 (Aalborg).png 100 (Aalborg).bb

clusternr: 100

1: aab-reserver spillede uafgjort.

2: aalborg: aab's superligareserver spillede tirsdag aften 1-1 i en testkamp mod kjellerup fra 2.

3: division.

101 (Aalborg).png 101 (Aalborg).bb

clusternr: 101

1: han vil egentlig bare gerne male.

2: ▫ esben hanefelt kristensen, ridefogedvej 7, aalborg, fylder fredag 65 år.

3: esben hanefelt kristensen samtaler næsten, som han maler.

Hansen 32 (Frederikshavn White Hawks, Frederikshavn).png
Hansen 32 (Frederikshavn White Hawks, Frederikshavn).bb

clusternr: 32

1: først med poesipark.

2: frederikshavn: med frederikshavns rå og uspolerede charme som guideline kunne byen fredag åbne landets første poesipark.

3: ideen til parken er helt efter norsk forbillede, da venskabsbyen larvik har udviklet sin park gennem de seneste ti år, men larvik har det fint med, at venskabsbyerne henter inspiration hos dem.

Hansen 39 (Frederikshavn).png Hansen 39 (Frederikshavn).bb

clusternr: 39

1: her får kontanthjælp en ny betydning.

2: frederikshavn: - jeg er sygemeldt og på kontanthjælp, så vi har ikke så meget at gøre med.

3: derfor søgte vi om at være med, siger anine sørensen, der onsdag sammen med sønnen waldemar for andet år i træk deltog i julefesten på restaurant frida i havnegade.

Hansen 54 (Socialdemokraterne, Skagen, Danmark, Los Angeles).png Hansen 54 (Socialdemokraterne, Skagen, Danmark, Los Angeles).bb

clusternr: 54*

1: ELLING: Sådan en augustdag ser Elling Å ganske fredsommelig ud, som den flyder stille forbi landsbyen Elling på sin vej mod udløbet i Kattegat lige nord for Rønnerhavnen.

2: Derfor havde Frederikshavn Kommune søndag inviteret transportminister Ole Birk Olesen (LA) til Elling for at besigtige åens løb under Skagensvej.

3: Men for at borgerne skal sove roligt i den lille by, skal hullet til Elling Å under Skagensvej laves større, så å-vandet ikke standses og presses tilbage.

Hansen 66 (Frederikshavn, Birgit Hansen, Socialdemokraterne, Hadsund, Venstre).png Hansen 66 (Frederikshavn, Birgit Hansen, Socialdemokraterne, Hadsund, Venstre).bb

clusternr: 66*

1: Rigtig god leder, der beskriver volden i folkeskolerne, og om hvordan det i dag er lærerne, der går rundt og har ondt i maven, når de skal på arbejde, hvor det førhen var eleverne, der havde ondt i maven, når de skulle i skole.

2: Denne oplevelse medførte, at jeg i en årerække var med i en rådgivningsgruppe i Allergiforeningens nordjyske afdeling, medlem af Allergiforeningens repræsentantskab - ud over at jeg i mit daglige arbejde med byggeri boligindretning m.

3: Nu er fortællingen om søhelten Peter Wessel Tordenskjold og hans huseren i området Fladstrand i starten af 1700-tallet blevet en vaskeægte "eksportvare", for i samarbejde med blandt andet Nationalmuseet kommer København i den kommende uge til at danne kulisse om en omgang Tordenskioldsdage i dronningens København.

Hansen 69 (Frederikshavn, Frederikshavn White Hawks, Fladstrand, Peter Wessel Tordenskiold).png Hansen 69 (Frederikshavn, Frederikshavn White Hawks, Fladstrand, Peter Wessel Tordenskiold).bb

clusternr: 69*

1: Nu er fortællingen om søhelten Peter Wessel Tordenskjold og hans huseren i området Fladstrand i starten af 1700-tallet blevet en vaskeægte "eksportvare", for i samarbejde med blandt andet Nationalmuseet kommer København i den kommende uge til at danne kulisse om en omgang Tordenskioldsdage i dronningens København.

2: Ikke mindst historien om Tordenskjold og alt det, han foretog sig for 300 år siden - i 1717 - ved Fladstrand, som var Frederikshavns navn frem til 1818, hvor byen blev købstad, er Jan Michael Madsen glad for.

3: Torsdag er Frederikshavns borgmester, Birgit Hansen, også til stede i København, og mon ikke hun - set i lyset af de aktuelle forhandlinger om et forsvarsforlig - vil benytte lejligheden til at slå på tromme for Frederikshavns århundredegamle historie som flådeby.

100 (Saeby).png 100 (Saeby).bb

clusternr: 100

1: jolle i knibe ud for frederikshavn sæby: lørdag aften blev sæby redningsstation alarmeret om, at en jolle med to mand om bord havde fået motorstop og var i knibe nær hirsholmene ud for frederikshavn.

2: situationen var farlig på grund af blæst og høje bølger, og bølgeskvulp medførte vand i jollen.

3:

199 (Frederikshavn Kommune).png 199 (Frederikshavn Kommune).bb

clusternr: 199

1: konservative helt på toppen.

2: det konservative folkeparti i frederikshavn kommune går til det kommende byrådsvalg til valg med sloganet "se muligheder" og vil arbejde for at gøre mulighederne til virkelighed - her illustreret med badehotellet på grenen.

3:

221 (Saeby).png 221 (Saeby).bb

clusternr: 221

1: bliv klogere på arveret og testamente.

2: sæby: nu har du mulighed for at blive afklaret omkring arvereglerne, og hvordan du opretter et testamente, når manegen og sæby bibliotek får besøg af to jurister fra kræftens bekæmpelse mandag 11.

3: september klokken 14.

502 (Frederikshavn Kommune).png 502 (Frederikshavn Kommune).bb

clusternr: 502

1: 64 procent offentligt ansatte i byrådet.

2: politik: i dag er 18 medlemmer af byrådet i frederikshavn kommune kommunalt eller offentligt ansatte.

3: det svarer til over 58 procent.

57 (Aalborg).png 57 (Aalborg).bb

clusternr: 57

1: varme kan foræres væk uden afgifter.

2: indland: dansk fjernvarme tror ikke, at en afgørelse fra skatteankenævnet om afgiftsfrihed for overskudsvarme fra facebooks nye server-center i odense får betydning for eksisterende aftaler om overskudsvarme fra virksomheder til fjernvarme.

3: dermed ser den altså ikke ud til at få konsekvenser for aalborgs fjernvarmeforbrugere, der i mange år har nydt godt af overskudsvarme fra aalborg portland.

178 (Aalborg).png 178 (Aalborg).bb

clusternr: 178

1: brandstationen i als skal selv klare de fleste opgaver.

2: als: for den lille brandstation i als betyder kontrakten mellem falck og nordjyllands beredskab, der trådte i kraft ved nytår, at det lokale mandskab så vidt muligt skal klare sig selv.

3: - som brandmænd vil vi i praksis mærke det på den måde, at når stationen i hadsund kommer for at hjælper ved større indsatser, vil de komme med lidt færre mænd.

212 (Holland, Amsterdam).png 212 (Holland, Amsterdam).bb

clusternr: 212*

1: Og pludselig kommer der en halvstor kanal-damper med et helt kreoler-orkester på dækket - vi havde godt hørt den nærme sig, men så holdt de en pause, og satte til gengæld i med trommer og basuner, lige da de kom rundt i svinget, hvor Hunni og jeg sad på kajkanten - Hunni var helt regulært ved at skide en grøn gris og forsøgte at stikke af, men jeg nåede at gribe hende.

2: For det første tror jeg nok, at bydelen i New York har udviklet sig en del siden 70'erne, og for det andet findes der et rigtigt Haarlem - en smuk og stolt by i Holland: Fra Amsterdam sejler du vestpå ad Noordzee Kanaal - en times tid eller halvanden, og du skal holde tungen lige i munden, for her møder man alle mulige mystiske skibe fra små ydmyge smadrekasser og motorbåde til store slæbebåde med mægtige pramme og alle tænkelige laster..

3: Folk jublede og klappede, men Hunni var skrækslagen, så jeg lempede hende ned ombord - ned under bordet i kahytten, hvor hun lå og rystede en hel time.

219 (Nyhedsbeauet AFP, Colombia).png 219 (Nyhedsbeauet AFP, Colombia).bb

clusternr: 219*

1: ITAGÜÍ: Indbyggerne i den colombianske by Itagüí fyldte forleden byens gader med madrasser og hængekøjer for at markere den årlige doven-dag, som er et forsøg på at opfordre stressede mennesker til at skrue ned for tempoet..

2: Indbyggerne i den colombianske by Itagüí fyldte forleden byens gader med madrasser og hængekøjer for at markere den årlige doven-dag, som er et forsøg på at opfordre stressede mennesker til at skrue ned for tempoet.

3: For 32 år siden fik beboeren Carlos Mario Montoya idéen om, ikke kun at fejre den travle industri, men også at fejre det langsomme liv.

243 (Morso Kommune, Socialdemokraterne).png 243 (Morso Kommune, Socialdemokraterne).bb

clusternr: 243*

1: Morsø Kommune lod i starten af 2016 Nykøbing Dag- og Industrirenovation ved vognmand Bjørn Filtenborg købe i første omgang syv hektar jord på sydsiden af Nørrebro i Nykøbing, så virksomheden i etaper kunne flytte fra en mere bynær placering og ud som genbo til kommunens egen genbrugsplads.

2: NYKØBING: Sagen om placeringen af virksomheden Nykøbing Dag- og Industrirenovation på det areal på Nørrebro i Nykøbing, der hidtil har været udlagt til lettere industri og håndværk, tager en ny drejning.

3: Hvis Jakob Kortbæk har den opfattelse, at Filtenborg burde have været placeret i Morsø Foodpark, er han en postgang for sent ude, siger Meiner Nørgaard.

2 (Nordjylland).png 2 (Nordjylland).bb

clusternr: 2

1: ældre trækker på beredskab.

2: nordjylland: når den store golfturnering made in denmark om få dage går i gang i gatten i vesthimmerland ventes der tusinder af tilskuere.

3: det øger sandsynligheden for, at der her kan opstå alvorlig sygdom eller der kan ske ulykker.

4 (Aalborg).png 4 (Aalborg).bb

clusternr: 4

1: lysets engel i aalborg.

2: teater.

3: "jomfru ane og maren turis – kærlighed, magt og hekseri".

12 (Socialdemokraterne, Aalborg Kommune).png 12 (Socialdemokraterne, Aalborg Kommune).bb

clusternr: 12

1: en anden prioritering.

2: kommunalvalg: i folketinget ønsker socialdemokraterne velfærd over skattelettelser.

3: i aalborg kommune ser de lidt anderledes på det.

17 (Aalborg).png 17 (Aalborg).bb

clusternr: 17

1: lette wienertoner i aalborg.

2: koncert.

3: aso og fire wiener- philharmonikere.

23 (Socialdemokraterne, Aalborg).png 23 (Socialdemokraterne, Aalborg).bb

clusternr: 23

1: salling får rooftop på 1400 kvm.

2: aalborg: der er knap gået et år, siden salling i aarhus slog dørene op til deres tagterrasse på toppen af stormagasinet, og den er blevet en bragende succes.

3: over 700.

Michael Johansen 0 (Aalborg Universitet, Per Michael Johansen).png Michael Johansen 0 (Aalborg Universitet, Per Michael Johansen).bb

---

clusternr: 0

1: medlem af nyt forenklingsudvalg.

2: udnævnt: louise gade, 45 år, direktør for via efter- og videreuddannelse, er udnævnt til at være medlem af regeringens nye forenklingsudvalg.

3: medlem af udvalget er også per michael johansen, professor og rektor for aalborg universitet.

Michael Johansen 1 (Nordjylland).png Michael Johansen 1 (Nordjylland).bb

clusternr: 1

1: - drop sparekrav på uddannelser.

2: nordjylland: uddannelse er altafgørende for at sikre, at der i fremtiden er job til nordjyderne og nordjyder nok til de job, der til den tid er.

3: og mange deltagere i den nordjyske åbningsdebat havde forslag og ideer til, hvad der bør gøres.

Michael Johansen 2 (Aalborg Universitet, Aalborg).png Michael Johansen 2 (Aalborg Universitet, Aalborg).bb

clusternr: 2

1: aau - bedst i europa inden for ingeniørvidenskab.

2: aalborg: aalborg universitet (aau) er det bedste universitet i europa inden for ingeniørvidenskab.

3: det vurderer u.

Michael Johansen 10 (Per Michael Johansen, Aalborg Universitet, Danmark, Aalborg).png Michael Johansen 10 (Per Michael Johansen, Aalborg Universitet, Danmark, Aalborg).bb

clusternr: 10*

1: To af de vigtigste talere bliver dels svenske Sverker Sörlin fra Kungliga Tekniska Högskolan i Stockholm - han fik i 1993 Nordens første professorat i miljøhistorie, dels australske Lynda Roper.

2: Konferencen finder sted i Aalborg Kongres & Kultur Center fra tirsdag til fredag i den kommende uge, og som generalsekretær for arrangementet har Poul Duedahl ikke haft meget tid til at kigge i gamle arkiver og den slags den seneste tid.

3: Programmet for de fire dage fylder adskillige sider og rummer omkring 200 arrangementer: Foredrag, diskussioner og som noget nyt også filmfremvisning om Første Verdenskrig med introduktion af Nils Arne Sørensen, der har skrevet bogen "Den Store Krig" om netop krigen i 1914-18.

Michael Johansen 11 (Aalborg Universitet, Per Michael Johansen, Danmark).png Michael Johansen 11 (Aalborg Universitet, Per Michael Johansen, Danmark).bb

---

clusternr: 11*

1: Samtidig, mener rektoren, at hvis Danmark og danske virksomheder skal sikre et fortsat rigt samfund, så betyder det, at der er behov for langt mere interaktion mellem samfund og universitet for at sikre, at den nyeste teknologi og den nyeste viden hele tiden er til rådighed, så Danmark kan sikre sin internationale konkurrenceevne og førerposition.

2: Hvis Danmark skal sikres som et fortsat rigt samfund og danske virksomheders globale konkurrenceevne skal styrkes, så kræver det langt mere interaktion mellem samfund og universitet.

3: Aalborg Universitet er dog godt klædt på til at klare de udfordringer, mener rektoren..

21 (Socialdemokraterne).png 21 (Socialdemokraterne).bb

clusternr: 21

1: klar til at lægge låg på airbnb.

2: københavn: et tysk par slæber kufferter op ad trappen den ene uge.

3: næste weekend er det en familie fra kina, der pusler rundt i naboens lejlighed.

48 (Venstre).png 48 (Venstre).bb

clusternr: 48

1: minister: ingen tegn på monopol hos tandlæger.

2: indland: et besøg hos tandlægen kan være en dyr fornøjelse, og nogle danskere vil sikkert ønske sig mere konkurrence blandt de danske tandlæger.

3: senest har det vakt bekymring, at kapitalfonde og andre storinvestorer er i gang med at købe sig ind i dansk tandpleje.

64 (Aalborg, Socialdemokraterne).png 64 (Aalborg, Socialdemokraterne).bb

clusternr: 64

1: min kandidat?.

2: valg: mit nuværende medlemskab af socialdemokratier er knyttet til en vælgerforening under vestre kreds i aalborg.

3: i en årrække var rasmus prehn vores folketingskandidat - med genvalg på genvalg.

260 (Aalborg, Venstre).png 260 (Aalborg, Venstre).bb

clusternr: 260

1: forslag splitter skoleudvalgg.

2: aalborg: ungebyrådets forslag om at indføre samfundsfag i 7.

3: klasse har splittet aalborgs ny skoleudvalg så meget, at det i stedet bliver op til byrådet at afgøre, hvordan det skal kunne lade sig gøre.

275 (Kelvin, Socialdemokratiet, Venstre).png 275 (Kelvin, Socialdemokratiet, Venstre).bb

clusternr: 275*

1: Hobro Kommune (2002-2007) stillede Jørgen Pontoppidan sig i spidsen for bestræbelser på at bygge et kommunalt plejehjem ved Hostrupkrogen i Hobros nordlige bydel..

2: S og K i byrådet indgik forlig om, at kommunen i stedet skulle udvide Hobro Alderdomshjem og Plejehjemmet Solgaven i Hobro..

3: I 2004 kom Jørgen Pontoppidan, borgmester i daværende Hobro Kommune, og hans parti Venstre, i mindretal i en sag om ældreboligbyggeri i Hobro..

Kastrup 22 (Aalborg, Mads Duedahl, Thomas Kastrup Larsen).png Kastrup 22 (Aalborg, Mads Duedahl, Thomas Kastrup Larsen).bb

clusternr: 22

1: spadestik til freja-hal.

2: aalborg: solen skinnede på vestby-klubben aalborg freja, da borgmester thomas kastrup-larsen, rådmand mads duedahl, frejas formand jan daucke sammen med de to unge frejanere katrine torp staffe og malthe kjølby tog første spadestik til den ny freja-hal.

3: hallen, der bliver en isoleret lethal, får en størrelse på 1570 kvadratmeter eller det, der svarer til 1,6 håndboldhal.

Kastrup 44 (Venstre).png Kastrup 44 (Venstre).bb

clusternr: 44

1: nødvendigt at tænke nyt.

2: satellitlægehuse: så er det kommunale budget endelig vedtaget for 2018, og venstre har i den forbindelse fået opbakning til, at kommunen gerne vil medfinansiere et nyt og moderne sundhedshus i storvorde.

3: det er vi glade for, men vi mangler stadig resten af den tidligere sejlflod kommune.

Kastrup 98 (Dansk Folkeparti, Thomas Krarup, Radikale Venstre, Aalborg, Socialistisk Folkeparti).png Kastrup 98 (Dansk Folkeparti, Thomas Krarup, Radikale Venstre, Aalborg, Socialistisk Folkeparti).bb

clusternr: 98*

1: Men Dansk Folkepartis spidskandidat Kristoffer Hjort Storm vil dog ikke give en garanti for, at DF peger på Tina French Nielsen som borgmesterkandidat, da NORDJYSKE Medier spørger om "det er helt sikkert, at DF peger på Tina French Nielsen som borgmester".

2: Udgangspunktet er at pege på Tina French Nielsen, men Kristoffer Hjort Storm vil ikke afvise at pege på Thomas Kastrup-Larsen (S).

3: Det var vi ikke, så derfor mistede vi et mandat på allersidste valgsted, siger Kristoffer Hjort Storm, der henviser til, at mandatet gik til Jørgen Hein (RV).

Kastrup 102 (Christiansborg, Aalborg Kommune, Venstre, Aalborg, Socialdemokraterne).png Kastrup 102 (Christiansborg, Aalborg Kommune, Venstre, Aalborg, Socialdemokraterne).bb

clusternr: 102*

1: På spørgsmålet om, om han som Venstre-politiker ikke i stedet for at kritisere byrådet i Aalborg, burde rette henvendelse til sine partifæller på Christiansborg, da det er Venstre, der leder den borgerlige regering og dermed er den mest oplagte kandidat til at finde finansiering til en Limfjordsforbindelse, placerer Kristian Andersen ansvaret hos oppositionen til regeringen, ikke mindst den socialdemokratiske leder Mette Frederiksen.

2: Det vil løfte kvaliteten i byen, styrke vækst og beskæftigelse og Aalborg Lufthavn, siger Kristian Andersen, der endvidere kritiserer det nuværende socialdemokratisk ledede byråd og i særdeleshed borgmester Thomas Kastrup Larsen for i for lang tid at have tøvet i forhold til det store anlægsprojekt.

3: at lægge maksimalt pres på Christiansborg for at få løst det her, siger borgmesteren, der mener, at Kristian Andersen i stedet burde rette skytset mod egne partifæller.

Kastrup 104 (Aalborg Havn, Aarhus, Royal Arctic Line, Aalborg, Thomas Kastrup-Larsen).png Kastrup 104 (Aalborg Havn, Aarhus, Royal Arctic Line, Aalborg, Thomas Kastrup-Larsen).bb

clusternr: 104*

1: Dagsordenen er givet på forhånd: At finde en løsning på den fastlåste konflikt mellem Aalborg Havn og RAL, der vil opgive Aalborg Havn som basishavn for skibstrafikken til og fra Grønland til fordel for havnen i Aarhus.

2: - Nu mødes vi endelig, og jeg ser frem til at høre, hvad Aalborg havn har at sige, og så håber jeg, at mødet er starten på en konstruktiv dialog om at finde en løsning, siger administrerende direktør Verner Hammeken fra RAL.

3: Mødet var det første mellem de to parter, siden RAL tilbage i juni meddelte, at rederiet vil opgive Aalborg som basishavn for skibstrafikken på Grønland til fordel for Aarhus med virkning fra og med 2019.