By Mushtaba Osmani (wgq323), Christan Lauersen (bfd962) & Mickey Johansson (mvx394)

Introduction

In this assignment, we were tasked with conducting simple data mining on a dataset consisting of transcripts of danish news broadcasts.
This collection of data contains transcripts from news broadcasts, the oldest being from 1939 and the newest from 2010. The transcripts have no punctuation and are very roughly split into different paragraphs, which supposedly should cover one story.
Because of the limited amount of data, combined with no easy way to sort through the different stories of each transcript, we decided to focus on the broadcasts themselves, instead of trying to query for a specific topic. We have tried to figure out how the broadcasts have changed over time.

The process of this portfolio assignment is admittedly still under progress and as a result, there may be some obvious shortcomings that we have not considered.

Pre-processing and Description of our Collection

First of all, a list of (Danish) stopwords has been used to exclude uninteresting words from the texts. Furthermore, all words were converted into lowercase. This was done in order to count all instances of words, for example. To make more relevant comparisons across the texts, which are divided into years spanning from 1939-2010, they were split into 4 different time periods. The older texts from 1939-1959, from 1960-1989 and the newer ones from 1990-1999 and 2000-2010. This opened up for the possibility of comparing how and if news stories have changed throughout the years. By using The collection consists of 25 different text documents with 227.872 characters in total. Meaning it is a small collection compared to other options.

Model and visualization of our topic

In an attempt to visualize the frequency of our words we used a simple graph highlighting the 10 most frequent words in each time period. The graph was only used to highlight the most frequent words excluding danish stopwords. The most interesting change the graphs visualized is that words in the first time periods were dominated by foreign countries and to lesser extent words related to war such as “styrker” (forces). The time period of the 90’s first included weather-related words such as “regn” (rain) and in the 2000s the most frequent words were now dominated by weather-related words. The most frequent words in the most recent broadcasts do not appear as frequently as the most frequent words in the older broadcasts does. As a result, it appears that the usage of words are generally more varied and grounded in more recent time periods. This is also reflected in terms of topic being more varied as we split by “/n”.

The average amount of paragraphs in the texts increases with time, meaning that the list with the oldest texts, have the fewest paragraphs, and the list with the newest texts, have the most paragraphs. This correlates with the average length of each paragraph.The list with the oldest texts have the longest paragraphs, and the paragraphs are shortest in the list with the most recent texts. This indicates that in the recent broadcasts more topics were covered in less detail, while the old broadcasts had fewer topics but they were covered in more detail.