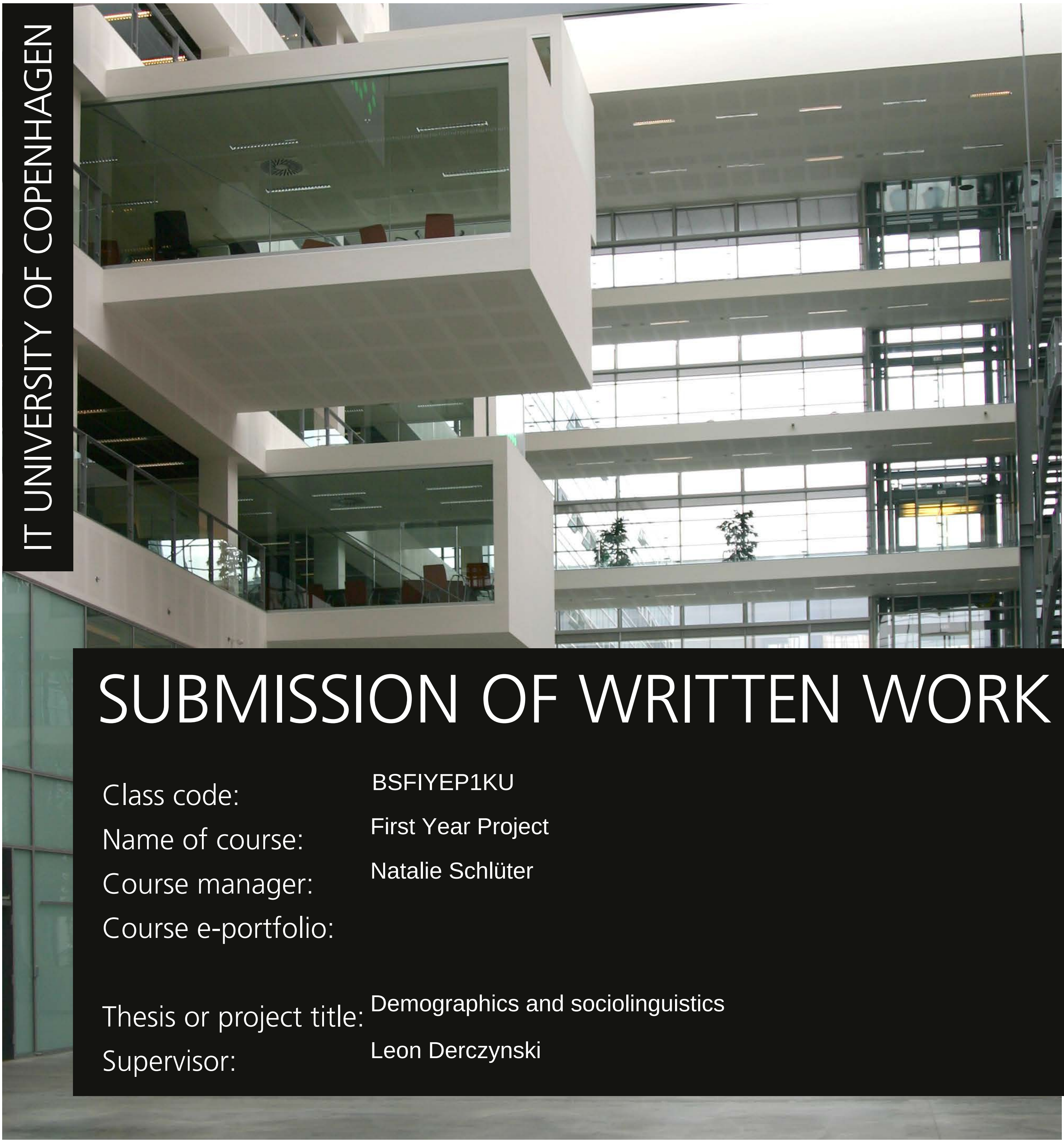


IT UNIVERSITY OF COPENHAGEN

SUBMISSION OF WRITTEN WORK

Class code: BSFIYEP1KU

Name of course: First Year Project

Course manager: Natalie Schlüter

Course e-portfolio:

Thesis or project title: Demographics and sociolinguistics

Supervisor: Leon Derczynski

	Full Name:	Birthdate (dd/mm-yyyy):	E-mail:
1.	Joachim Moe Osterhammel	07/12-1995	joos@itu.dk
2.	Stefan Maksimovic	13/01-1999	stma@itu.dk
3.	Emil Dichmann	04/11-1998	emdi@itu.dk
4.	Cindy Weronika Wendowski	25/12-1994	ciwe@itu.dk
5.	Gergő Koncz	06/08-1995	geko@itu.dk
6.			@itu.dk
7.			@itu.dk
8.			@itu.dk

Introduction

The president of the United States of America is arguably the most powerful and impactful man on Earth. Campaigns, elections and speeches all reach a world-wide audience. In this paper, we wish to analyze the most important political tool, *speeches*. Specifically, we wish to investigate the major differences between Republican and Democratic presidents, and how well we can classify a given speech as either Republican or Democratic. This paper also investigates other metrics such as wartime, historical periods and extramarital affairs.

Methodology

In order to analyze the textual data, we utilized the python modules SpaCy as well as textStats. SpaCy cleaned and tokenized all of the speeches, as well as pos-tagging them. In analyzing the difficulty of readability, we employed two measures, `Flesch Kincaid`[1] and `SMOG`. The FK method uses the following formula $206.835 - 1.015 \cdot (\frac{\text{total words}}{\text{total sentences}}) - 84.6 \cdot (\frac{\text{total syllabes}}{\text{total words}})$. It outputs a score between 0 and 100, where 60 indicates a readability difficulty of *New York Daily News* and a score between 0 and 30 indicates a readability difficulty of a university graduate.

The `SMOG` index uses the following formula $= 1.0430\sqrt{\text{number of polysyllables } \times \frac{30}{\text{number of sentences}}} + 3.1291$ and yields which school grade difficulty the text corresponds to. For example, Trump's speeches are on average classified as $6th$ grade reading.

This answered the questions regarding the development, or lack of it, in the difficulty of American presidents' speeches. In order to classify the speeches as Republican or Democratic, wartime or peace-time, we employed a binary logistic regression model. To predict the historical period in which the presidents held the administration, we used Multinomial Bayes Classifier[2].

The binary logistic regression is an extension of simple linear regression where the target variable we wish to predict is not continuous but binary, whereas MBC is based on Bayes theorem.

The reasoning behind avoiding measures such as `n-grams` and `TF-IDF` in the analysis, is that all presidents share customary sentences, such as *"God Bless America"*. This required expanding to quad-grams or penta-grams, where results were still flawed and skewed by these commonalities. The reasoning behind not using LDA is that our corpus size per president was too small. Presidents speak somewhat rarely, which in return means that the probability of error (probability of wrongful statistical inference) made LDA too volatile.

[1] FK henceforth

[2] MBC henceforth

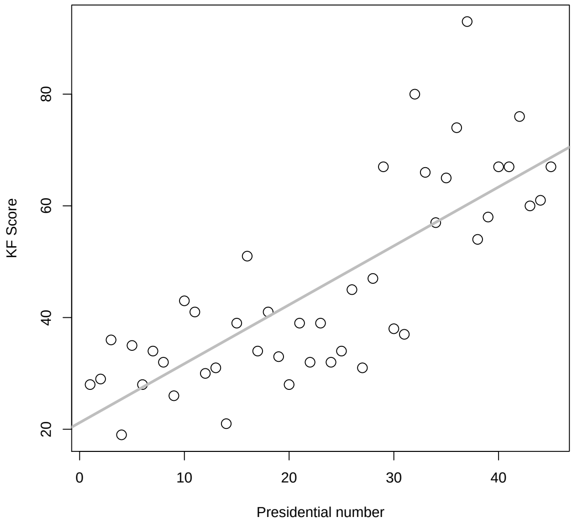
Data

The data was collected from Millercenter who holds copies and official transcripts of most of the public presidential speeches and letters. A total of 989 texts were scraped, which varies drastically in length. A scraper was used to collect the links from the overview page, and then scrape the transcripts from each of the respective speech pages. The speeches were converted directly to text elements and had HTML tags and elements removed. The transcripts were indexed by speaker and date.

Results

In our analysis of readability difficulty, we found a positive correlation between time (President number) and a decrease in reading difficulty (KF-score), depicted in Figure 1. The graph depicts a significant decrease in readability difficulty, with a noticeable leap around the $1920's$ to $1930's$. In Table 1, all prediction scores for the four categories are reported. We ended up with four different metrics with prediction scores, ranging from 57% to 84%. Other models, such as Naïve Bayes and K-Nearest Neighbour Classifiers were trialed but yielded overall lower prediction scores. The multinomial Bayes Classifier prediction scores are averaged for the following historical eras: Before and after The Civil War, World War*s* and the Cold War. They are averaged because we wish to predict periods in plural and not a singular period.

Figure 1: Readability over time



	Democrat	Cheater	Wartime	Historical period
Accuracy	0.67	0.69	0.72	0.66
Precision	0.70	0.64	0.72	0.62*
Recall	0.79	0.57	0.84	0.60*
F1	0.74	0.61	0.77	0.59*

*Table 1: Precision scores for the four categories (*average of scores for five historical periods).*

Interpretation

The results clearly point to evidence that the readability difficulty has decreased over time. A closer analysis shows that Democratic presidents have relatively more difficult speeches than Republicans. This is also supported by evidence that Republicans use more punctuation and have fewer unique sentences. This indicates that Republicans having shorter and more repeated sentences. However, there is a general trend from both parties of speaking in a more digestable manner.

In order to classify the speeches by party, we found that Democrats and Republicans have significant differences in their usage of words. An example of this, is that a frequent usage of the personal pronoun ”we”, significantly indicates the speech as being Democratic, whereas a more frequent usage of the word ”America” signifies a speech as being Republican. This somewhat reflects the ideological standpoint of both parties as expected.

Error Analysis

In the interpretation of results, one main question stands – is it a fair comparison? When analyzing time series data, a heap of contextual factors have to be taken into account. There might be shifts in the political landscape, such as changes in voting rights which influences the strategical aspect of presidential speeches. Other major societal shifts such as technological development, media norm changes also heavily influence our results. As example, we see a steep decrease of reading difficulty around the first world war, where also the radio became a household commodity.

Even though our logistic model predicts with an accuracy score of $\sim$ 70%, the reason behind not having a better score might be that presidents speak to the median voter, and not to a party specific audience. With all of this in mind, we ask the reader not to interpret these findings as universal facts, but more as preliminary research, as time constraints did not allow us to take these factors into account.

Concluding Remarks

In this paper we set out to investigate the major differences between Republican and Democratic presidents. We found that differences were both in the readability of the respective speeches, and differences in usage of words. In classifying speeches, we achieved an accuracy of 70%, which might reflect changes in the usage of speeches as a political tool.