1 slajd:

Welcome all to our presentation. Today we will take you to the trip throughout the centuries of English language evolution.

2 slajd:

The goal of this project is to show you how English language evolved between 17th to 20th century. We conducted the sentiment, part of speech and some additional analysis of literature to capture some interesting language tendencies.

3 slajd:

The books were collected from Guttenberg Project. To somehow simplify the analysis we used rank of best books for century from goodreads.com to collect books based on the list there. With that approach we encountered the issue in case of books from 20th century, because most of them were not present at all due to copyrights. To solve this problem we grabbed every first available book. Generally we the number of books we collected for each century was equal to 15

4 slajd:

Having books we had to somehow make them usable as the input data for our analysis. As I’m a big fan of python language tools most of the linguistic operations was done in this language. We used

* Regex for words extraction
* Stanford NLP in Python for Lemmatization and Part of Speech
* Sentiment Analysis with Vader in NLTK Python
* Also we got spatial data with coordinates of birthplaces of authors with Openstreetmap API

**5 slajd pusty**

6 slajd:

Kasia

7 slajd:

Kasia

8 slajd:

The eighteenth century is a clear winner in the competition of total numer of words. It is also numer one in the case of number of unique words. However what has to be said is that it might be the specific of our data. We did not have chances to look at the books at some global level. So our dataset is highly limited and not really representative in this aspect.

9 slajd:

In the area of proportion of unique words. It is highly correlated with the previous plot. We can observe that the best proportion of unique words to all have these centuries where there were the lowest number of words. This actually captures that number of unique words in language is limited. Therefore after some point the higher number of all words will not indicate the higher number of unique words.

**10 slajd pusty**

11 slajd:

To see how language evolved in case of part of speech we analyzed proportions of 4 most important speech parts

* Adjective
* Adverb
* Noun
* Verb

and group of the rest as others. We can observe that the most similar in the area of distribution of Parts of Speech are 17th and 20th vs 18th and 19th. We could conclude that the 20th century is the most expressive one because the percentage of adjectives. Also there is also the highest percentage of nouns and verbs. It is obviously the question whether this is a real pattern or just some data specific thing. It might be the case that the 20th century words were the ones most understandable for the POS algorithm. Therefore it achieved the best performance in classifying them.

**12 slajd pusty**

13 slajd w górę kasia