**Introduction**

Can the language used in an article, blog post or book give us any clues about their publication date?

In this project, we develop a supervised learning algorithm that reveals and learns the correlation between the content of an article and its publication date. The result is used to date articles without date-stamp. E.g., assign a date to news spinets without reference, categorize articles according to their relevant time periods. The motivation is the irregular use of certain words as seen in news articles and Google analytics.

**Methodology**

Having access to the New York Times API, collected the content of a 1000 articles/month for 30 years (1987 to 2017). First we tokenize, stem and filter words that would not be helpful features. The final step in data collection was filtering out articles with too little content (such as video or slide-show posts).

The main step before learning is feature selection, i.e. choosing words whose frequency reveals specific events in time. We then iterate between feature selection and learning using SVM, Logistic Regression, Deep Learning, and Naive Bayes. We did the learning on both a monthly and yearly basis.

**Feature Selection**

Due to the large number of features (or words) a heuristic method for feature selection is used. A measure of fluctuation or average velocity of word appearance is given by:

$$score_w(words) = \sum_{y=1987}^{2016} f_w(y+1) - f_w(y)$$

where $f_w(y)$ the frequency of the word $w$ in year $y$, and proportionally constant $C_w$ gives more weight to frequent words.

Feature selection is a challenge when the corpus contains around 300K words. More features causes over fitting while less causes bias.

**Results**

We study the error of our model as a function of size of the training set for Naive Bayes classifier. It can be seen that with the current size of training set, both 1-year prediction error and standard deviation of error have reached their asymptotic values.

**Conclusion & Future Work**

We showed that there is a correlation between the content of news articles and their publication date. Naive Bayes did better than NN and Softmax. Feature selection is the biggest challenge. In the future We would like to:

- Use mutual information and for feature selection
- Word2Vec and k-means for studying the dynamics of meaning and for reducing the number of features.
- Using n-grams and focusing more attention on names.
- Study how the error improves with the time range and across source.

**Models Evaluation**

Our preliminary results are given by the following table. Better ones are coming as we speak!

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy Training Set</th>
<th>Accuracy Dev. Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB-sklearn</td>
<td>35%</td>
<td>24%</td>
</tr>
<tr>
<td>NB-MAT</td>
<td>31%</td>
<td>29%</td>
</tr>
<tr>
<td>LinR-MAT</td>
<td>21%</td>
<td>9%</td>
</tr>
<tr>
<td>LR</td>
<td>40%</td>
<td>22%</td>
</tr>
<tr>
<td>NN-SM</td>
<td>22.4%</td>
<td>11.8%</td>
</tr>
<tr>
<td>NN-LinR</td>
<td>5.6%</td>
<td>5.5%</td>
</tr>
</tbody>
</table>

The accuracy of the NN with linear regression is low but provides information for the correct prediction vicinity. This information will be modeled as a Gaussian pdf and with the combination of Softmax prediction the accuracy is boosted.