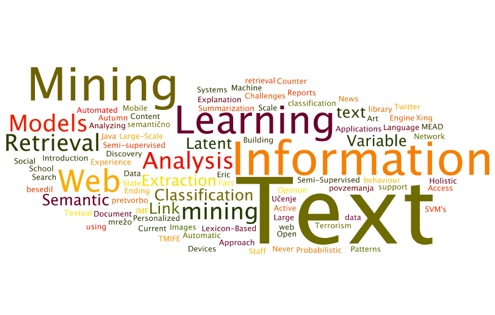
**Text Analytics**

**Exercise 1 Report**

**Papastamos Konstantinos**

**Kanellis Georgios**

**Tsolas Leonidas**



# Introduction

The exercise given is described below:

**1.** Implement (in any programming language) a bigram and a trigram language model for word sequences (e.g., sentences), using Laplace smoothing (slide 8) or optionally (much better in practice) Kneser-Ney smoothing (slide 38). Train your models on a training subset of a corpus (e.g., from the English part of Europarl). Include in the vocabulary only words that occur, e.g., at least 10 times in the training subset; use the same vocabulary in the bigram and trigram models. Replace all out-of-vocabular (OOV) words (in the training, development, test subsets) by a special token \*UNK\*. Assume that each sentence starts with the pseudo-token \*start\* (or the pseudo-tokens \*start1\*, \*start2\* for the trigram model) and ends with \*end\*.

**2.** Check the log-probabilities that your trained models return when given (correct) sentences from the test subset vs. (incorrect) sentences of the same length (in words) consisting of randomly selected vocabulary words.

**3.** Estimate the language cross-entropy and perplexity of your models on the test subset of the corpus, treating the entire test subset as a single sequence, with \*start\* (or \*start1\*, \*start2\*) at the beginning of each sentence, and \*end\* at the end of each sentence. Do not include probabilities of the form P(\*start\*|…) (or P(\*start1\*|…) or P(\*start2\*|…)) in the computation of perplexity, but include probabilities of the form P(\*end\*|…).

**4.** Optionally combine your two models using linear interpolation (slide 10) and check if the combined model performs better.

For our implementation we used python and the nltk package. Below all steps taken will be described in detail along with the necessary result screenshots.

# Data Loading

First the source release file of the europarl corpus, as found [here](http://www.statmt.org/europarl/), was downloaded. Only the english folder (“en”) was used to load the data. After the files were ready, we iterated over each file and