# LINGI2263 Computational linguistics
# Assignment 2 : Text Categorization

gr10 : Mulders Corentin, Pelsser Francois
29/03/2012

# 2 Preprocessing

To tokenize the files we had to determine where to split the sentences in words and wich types of expressions should be replaced by a token representing the type.
For slitting the tokens we simply used the whitespace as a separator as well as special characters that we didn't include in our types definitions. With the exception of the special characters ”-” and ”’” wich can appear inside some tokens that we didn't wish to split.
We also added markers for start and end of sentences. The start of sentences where placed before capitalized words followed by any sequence of tokens finishing by a dot or a newline. The end of sentences where then placed after any dot followed by either a start of sentence, a newline or the end of the string.

## 2.1 Mapping expressions to types

We replaced some types of expressions by aliases :

**email** replaces any email

**date** replaces the dates

**kikoo** replaces words such as ”xoXOxo” or any variation.

**smiley** replaces any smiley

**math** replaces mathematical expressions

**punctuationfreak** replaces multiple punctuation marks.

**repeatedchars** replaces any word containing a letter repeated 3 or more times in a row.

**weirdcaps** replaces words with a least a non capitalized character followed by a capitalized one.

We also replaced occurences of ”’s” by ”is” and of ”’m” by ”am”. However for the negations such as in ”don't” we decided to keep the ”’t” attached to ”don” instead of replacing it all by ”do not”.

## 2.2 Most frequent tokens

After tokenizing the corpus (composed of the training files for both male and female blogs) the total number of tokens retrieved in the lexicon is 52505 distinct tokens.
Here are the top 20 most frequent tokens types extracted from the corpus. We removed the markers ¡s¿ and ¡/s¿ used to represent respectively the beginning and the end of a sentence from this top list. They were in the top 2 postiions and weren't really interesting.

| token | frenquency |
|---|---|
| the | 34631 |
| to | 22906 |
| and | 21281 |
| a | 19722 |
| of | 18168 |
| I | 16620 |
| is | 16436 |
| in | 11903 |
| that | 9428 |
| for | 7512 |
| it | 7179 |
| on | 6184 |
| my | 5854 |
| was | 5843 |
| with | 5705 |
| you | 5656 |
| have | 4692 |
| this | 4468 |
| be | 4186 |
| as | 4157 |

We can see that the word "the" is the most frequent one as expected.

# 3 Word and N-gram counts

Here is the graph of the number of unigrams, bigrams and trigrams for each frequency :

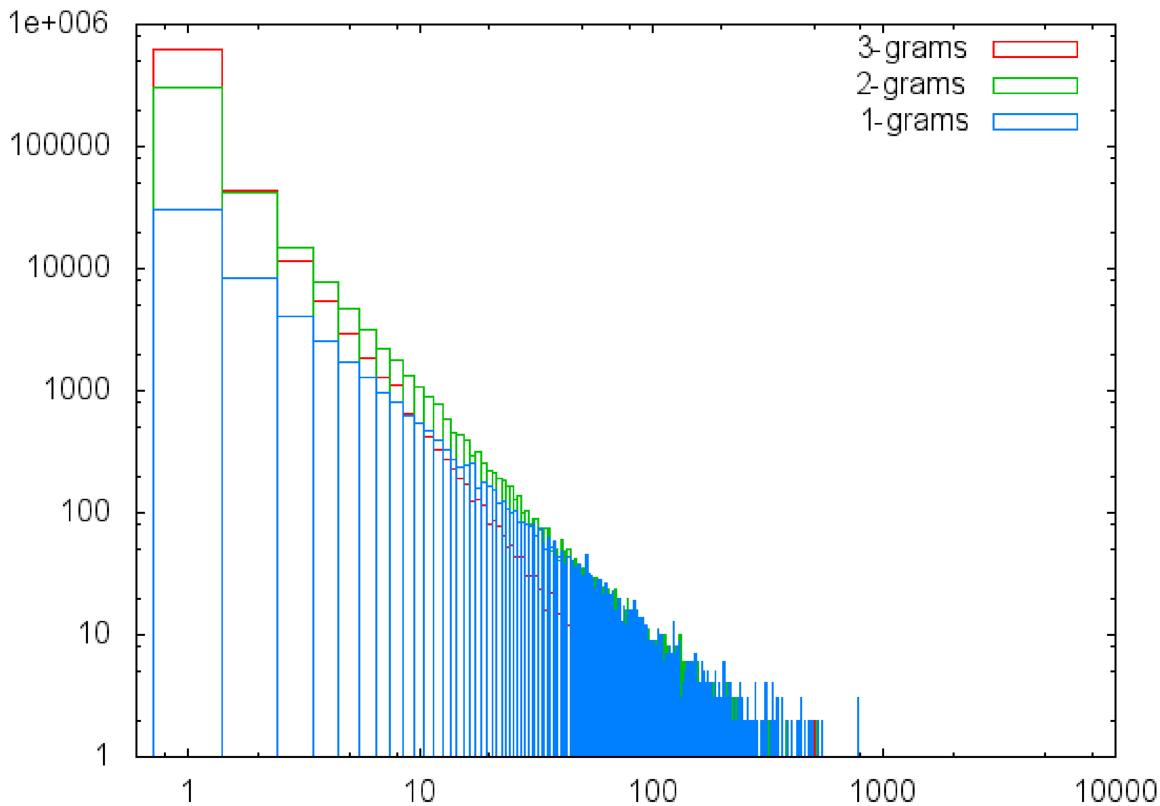
1e+006
100000
10000
1000
100
10
1
1
10
100
1000
10000
3-grams
2-grams
1-grams

**Figure 1**

We can see the unigrams histogram in blue and it is linear in log-log scale so Zipf's law is respected.
The same seems to be true for bigrams and trigrams.

# 4 N-gram estimation

We computed the mean perplexity over each test set using each training set for n-grams with $n \in [1, 5]$ and with the basic laplace add one smoothing and the linear interpolation smoothing.
We didn't manage to keep the consistency for the laplace model so the associated perplexities are a little bit crazy...

## 4.1 1-gram with laplace smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 140.07300098438208 | 0.03219323101098795 | 165.55490835053433 | 0.03644707958261017 |
| male | 134.49432287066887 | 0.03219323101098795 | 143.43047078221815 | 0.03644707958261017 |

## 4.2 1-gram with linear smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 1181.4392640498015 | 0.03219323101098795 | 1396.3652360335718 | 0.03644707958261017 |
| male | 1273.3247776421954 | 0.03219323101098795 | 1357.9277431026294 | 0.03644707958261017 |

## 4.3 2-gram with laplace smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 6597.051090993472 | 0.03219323101098795 | 7526.464184988605 | 0.03644707958261017 |
| male | 6751.450681527784 | 0.03219323101098795 | 7068.504328026753 | 0.03644707958261017 |

## 4.4 2-gram with linear smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 85.85618860497664 | 0.03219323101098795 | 98.29094897820075 | 0.03644707958261017 |
| male | 94.7289894781771 | 0.03219323101098795 | 93.48257298944773 | 0.03644707958261017 |

## 4.5 3-gram with laplace smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 19419.704289536854 | 0.03219323101098795 | 21080.08273975077 | 0.03644707958261017 |
| male | 19857.104934164177 | 0.03219323101098795 | 20674.655403927376 | 0.03644707958261017 |

## 4.6 3-gram with linear smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 74.05775785805339 | 0.03219323101098795 | 96.82676770775483 | 0.03644707958261017 |
| male | 87.29893378787449 | 0.03219323101098795 | 88.4301357336149 | 0.03644707958261017 |

### 4.7 4-gram with laplace smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 23270.25113767307 | 0.03219323101098795 | 24783.006381785264 | 0.03644707958261017 |
| male | 23512.386931883306 | 0.03219323101098795 | 24376.779729652386 | 0.03644707958261017 |

### 4.8 4-gram with linear smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 132.88670099907827 | 0.03219323101098795 | 174.63117442256916 | 0.03644707958261017 |
| male | 151.2064322011711 | 0.03219323101098795 | 153.94702577858564 | 0.03644707958261017 |

### 4.9 5-gram with laplace smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 23428.567965680333 | 0.03219323101098795 | 24823.951627660903 | 0.03644707958261017 |
| male | 23588.180660734663 | 0.03219323101098795 | 24415.93450835315 | 0.03644707958261017 |

### 4.10 5-gram with linear smoothing :

| | On female test set | | On male test set | |
|---|---|---|---|---|
| training set | mean perplexity | mean perplexity OOV rate | mean perplexity | mean perplexity OOV rate |
| female | 164.70561612766298 | 0.03219323101098795 | 214.29336077008773 | 0.03644707958261017 |
| male | 190.28328162528635 | 0.03219323101098795 | 193.30692013526775 | 0.03644707958261017 |

### 4.11 Conclusions from the perplexities

We notice that most of the time the perplexity islower on the same test set as the training set used. This is expected since sets from the same gender match better.

We also notice that the lowest perplexity is for linear smoothing with 3-grams. We'll see later if this reflects on the quality of the prediction.

## 5 Categorization of blog messages per gender

Here are the confuction matrices that we obtained with unigrams, trigrams and pentagrams while using laplace add one smoothing and linear interpolation smoothing. Each line corresponds to the results obtained for the lines of a given test set. Columns represent the value guessed.

### 5.1 1-grams with laplace smoothing :

| | male | female |
|---|---|---|
| male | 596 | 78 |
| female | 342 | 205 |

### 5.2 1-grams with linear smoothing :

| | male | female |
|---|---|---|
| male | 326 | 348 |
| female | 93 | 454 |

## 5.3 3-grams with laplace smoothing :

| | male | female |
|---|---|---|
| male | 382 | 292 |
| female | 151 | 396 |

## 5.4 3-grams with linear smoothing :

| | male | female |
|---|---|---|
| male | 402 | 272 |
| female | 148 | 399 |

## 5.5 5-grams with laplace smoothing :

| | male | female |
|---|---|---|
| male | 411 | 263 |
| female | 198 | 349 |

## 5.6 5-grams with linear smoothing :

| | male | female |
|---|---|---|
| male | 399 | 275 |
| female | 164 | 383 |

## 5.7 Conclusions

### 5.7.1 Optimal model order and smoothing technique

First of all we can easilty tell that the linear interpolation smoothing tends to be better than the laplace add one. For unigrams laplace results on detecting female test samples are disastrous.
Then about the optimal order, unigrams be be excluded since the results are pretty random. However with linear smoothing the results with 3-grams and 5-grams are nearly the sames. But since 5-grams require more computation power the 3-grams can be selected as optimal.

### 5.7.2 Correlation to the perplexity results

We noted that the lowest perplexities were obtained with 3-grams linear smoothing. This is also the model order and smoothing technique that we selected as optimal from the confusion matrices so the perplexity results were a good indicator of the classification accuracy.