Récapitulatif

N-Gram-Based Text Categorization

->System is based on calculating and comparing profiles of N-gram frequencies

### Zipf’s Law

The nth most common word in a human language text occurs with a frequency inversely proportional to n.

-> Human languages invariably have some words which occur more frequently than others.

Use the N-gram frequency profile technique to classify documents according to their language without building a lexicon or a set of morphological processing rules. Instead, we need merely obtain modestly sized sample texts (10K to 20K bytes), calculate the N-gram frequency profiles, and use those to classify the documents.

### Generating N-Gram Frequency Profiles

1. Tokenization and preprocessing

Split the text into separate tokens consisting only of letters and apostrophes. Digits and punctuation are discarded. Pad the token with sufficient blanks before and after.

2. Generation of N-grams

Scan down each token, generating all possible N-grams, for N=1 to 5. Use positions that span the padding blanks, as well.

3. Counting N-grams

Hash into a table to find the counter for the N-gram, and increment it. The hash table uses a conventional collision handling mechanism to ensure that each N-gram gets its own counter.

4. Sorting N-grams by frequency

Sort those counts into reverse order by the number of occurrences. Keep just the N-grams themselves, which are now in reverse order of frequency.

-> The resulting file… category samples

● Observations

* The top 300 or so N-grams are almost always highly correlated to the language = differs from languages
* The very highest ranking N-grams are mostly uni-grams (N=1), and simply reflect the distribution of the letters of the alphabet in the document’s language. After that come N-grams that comprise function words (such as determiners) and very frequent prefixes and suffixes. There is, of course, a long tail to the distribution of language-specific N-grams, and it goes well past 300.
* Starting around rank 300 or so, an N-gram frequency profile begins to show N-grams that are more specific to the subject of the document. These represent terms and stems that occur very frequently in documents about the subject.
* There is nothing special about rank 300 itself, since Zipf’s law gives us in fact a very smooth distribution curve. Rather, we arrived at this number mostly by inspection. Doubtless, one could do more elaborate statistics and choose an optimal cutoff rank for a particular application.

### Comparing and Ranking N-Gram Frequency Profiles

- Take two N-gram profiles and calculates a simple rank-order statistic we call the “out-of-place” measure. This measure determines how far out of place an N-gram in one profile is from its place in the other profile.

ex.

the N-gram “ING” at rank 2 in the document, but at rank 5 in the category -> 3 ranks out of place

-The sum of all of the out-of-place values for all N-grams is the distance measure for the document from the category.

- Pick the category with the smallest distance.