Subject Technology Features

This document focuses on features related to the distribution of a technology term. There are four classes of those features, we will look at each class and give descriptions and some informal formulas.

Class 1: title and beginning of document features

The first class is for binary features that encode whether a term or part of the term occurs in the title or the beginning of the document.

feature	description
in_title	term occurs in the title of the document
in_title_h	... same, but now for the head of the term
in_title_un	... same, but now for any unigram in the term
in_title_bi	... same, but now for any bigram in the term
in_beginning	term occurs in the beginning of the document
in_beginning_h	... same, but now for the head of the term
in_beginning_un	... same, but now for any unigram in the term
in_beginning_bi	... same, but now for any bigram in the term

The feature value is `true` or `false`. A term occurs in the beginning if it is in the first 10% of terms listed in order. We should pay around with this setting, it may for example be better to define this as occurring amongst the first 20 terms, especially for larger documents.

Class 2: frequency features

Frequencies are relative frequencies on a scale from 0 to 100, they are measured relative to the number of terms in a document.

feature	description
freq	frequency of the term in the document
freq_h	... same, but now for the head of the term
freq_un	... same, but now for any unigram of the term
freq_bi	... same, but now for any bigram of the term

Let's take "principal component analysis" as an example. If this term occurs 10 times in a document with 200 terms then `freq=5`. Now let's say that there are 20 other terms in the document that include "principal", "component" or "analysis". Then `freq_u=15` (20 other terms plus the 10 occurences of the term itself, divided by 200).

Class 3: relative position and range features

These are all numbers from 0 to 100.

feature	description
relpos	overal relative position of a term in a document
relpos_h	... same, but now for the head of the term
relpos_un	... same, but now for all unigrams of the term
relpos_bi	... same, but now for all bigrams of the term
range	range over the document where the term occurs
range_h	... same, but now for the head of the term
range_un	... same, but now for all unigrams of the term
range_bi	... same, but now for all bigrams of the term

The relative position is an average of the position of all term instances in a document. If a term occurs twice, once at the begining and once right in the middle, then `relpos=25.5` (that is, 1 plus 50, divided by 2). For the other relative position features you also include occurrences in other terms. For the range, if a term has only one instance in a document then `range=0`, if it occurs at the begining and right in the middle then we have `range=49` (that is, 50 minus 1).

Here are some definitions:

notation	definition
term = { $inst_{p_1}$ ... $inst_{p_n}$ }	all instances of a term type, ranging from position p_1 through p_n
$inst_p$	the term instance at position p (position is in the list of all instances)
T	all instances of all term types
n = | term |	number of instances of *term*
N = | T |	total number of all term instances

Some formulas for the features:

feature	formula
relpos(t)	$\sum_{t_p \in t}(p/n)$
range(t)	p_n - p_1

The others are somewhat harder to express in a neat formula, but they are conceptually not hard. For example, to get *relpos_h(t)*, you first do the following:

1. h_t = the head of term t
2. τ = the set of term instances that h_t occurs in, that is { t_i \| h_t in t_i for all $t_i \in T$ }

And now you can do something similar to *relpos(t)*: relpos_h(t) = $\sum_{t_p \in \tau}(p/|\tau|)$

So first you gather the instances and then calculate the relative position. This works in a similar way for the range features.

Class 4: shared content features

Finally, there is a set of features that measures shared content of the term with other terms. All numbers are again relative numbers on a scale from 0 to 100.

feature	description
fan_out	how often does the term occur in any other term
fan_out_h	... same, but now for the head word
fan_out_u	... same, but now for any unigram in the term
fan_out_bi	... same, but now for any bigram in the term
fan_in	how often do other terms occur in term
fan_in_h	... same, but now for the head word
fan_in_un	... same, but now for any unigram in the term
fan_in_bi	... same, but now for any bigram in the term

Let's take again the example of 10 occurences of "principal component analysis" in a document with 200 terms and let's say that there is also a term "principle components" with 20 occurrences. Then we will have `fan_out=5` for "principle components" because it occurs in 10 other terms. And we will have `fan_in=10` for "principle component analysis" because there are 20 terms that are contained in it. If there were also to be 10 instances of "principle components approach" then we would get `fan_out=10` for "principle components".

The story gets a little bit more complex for the variants for head word, unigrams and bigram, but those complications are similar to theones outlined for the class 3 features.