Assignment 2 - Words and The Company They Keep

Adrián Pauer

1 Best Friends

In this part we computed pointwise mutual information for bigrams and distant bigrams using formula

$$I(a,b) = \log_2\left(\frac{p(a,b)}{p(a)p(b)}\right).$$

To estimate the bigram and unigram distributions, we consider the first 8000 words that appear at least 10 times in the dataset.

Our results show that some word pairs exhibit negative PMI values. Table 1 presents examples of such pairs extracted from the Czech dataset.

Table 1: Pairs with negative PMI

pair	PMI
(’začala’, ’,’)	-0.310172
(’,’, ’členů’)	-0.688684
(’,’, ’v’)	-0.369542
(’LN’, ’.’)	-1.105118
(’a’, ’že’)	-0.008790
(’.’, ’a’)	-2.119988
(’.’, ’letech’)	-0.193345
(’.’, ’E’)	-0.419227
(’”’, ’v’)	-0.865643
(’komise’, ’a’)	-0.275223

In general, a bigram obtains a negative PMI value when $\frac{p(a,b)}{p(a)p(b)} < 1$ or equivalently, when $p(a,b) < p(a)p(b)$. This indicates that the two words are more likely to appear separately in the text rather than together as a pair. In other words, the words ”dislike” each other, leading to a negative pointwise mutual information value.

1.1 Bigrams

We obtained results for both the Czech and English datasets, which are presented in Tables 2 and 3.

Table 2: PMI for czech dataset and bigrams

pair	PMI
(Hamburger, SV)	14.288950
(Los, Angeles)	14.062442
(Johna, Newcomba)	13.762882
(Č., Budějovice)	13.633599
(série, ATP)	13.468968
(turnajové, série)	13.434411
(Tomáš, Ježek)	13.428981
(Lidové, noviny)	13.329922
(Lidových, novin)	13.271028
(veřejného, mínění)	13.062442
(teplota, minus)	12.981522
(Ján, Čarnogurský)	12.955527
(jaderné, zbraně)	12.955527
(Milan, Máčala)	12.897811
(lidských, práv)	12.862877
(společném, státě)	12.708434
(akciových, společností)	12.692492
(Pohár, UEFA)	12.625378
(privatizačních, projektů)	12.615677
(George, Bushe)	12.603010

Table 3: PMI for english dataset and bigrams

pair	PMI
(La, Plata)	14.169370
(Asa, Gray)	14.031867
(Fritz, Muller)	13.362016
(worth, while)	13.332869
(faced, tumbler)	13.262480
(lowly, organised)	13.216899
(Malay, Archipelago)	13.110477
(shoulder, stripe)	13.053893
(Great, Britain)	12.914557
(United, States)	12.847442
(English, carrier)	12.525514
(specially, endowed)	12.401817
(Sir, J)	12.377364
(branched, off)	12.377364
(mental, qualities)	12.362016
(de, Candolle)	12.362016
(Galapagos, Archipelago)	12.344942
(red, clover)	12.323880
(self, fertilisation)	12.316928
(systematic, affinity)	12.251833

1.2 Distant bigrams

For distant bigrams, we consider both directional orders and distance of 50 words. We obtained the following results.

Table 4: PMI for Czech dataset and **distant** bigrams

pair	PMI
(výher, výher)	9.855555
(žel, žel)	9.136365
(13h, 13h)	8.855555
(Sandžaku, Sandžaku)	8.826409
(Petrof, Petrof)	8.785767
(Bělehrad, Benfica)	8.688906
(Benfica, Bělehrad)	8.688906
(CIA, CIA)	8.504481
(IFS, IFS)	8.437844
(Benfica, Atény)	8.340982
(13h, zataženo)	8.340982
(Atény, Benfica)	8.340982
(zataženo, 13h)	8.340982
(13h, skoro)	8.322367
(skoro, 13h)	8.322367
(39, 39)	8.300340
(Atény, Bělehrad)	8.266982
(IV, výher)	8.266982
(výher, IV)	8.266982
(Bělehrad, Atény)	8.266982

Table 5: PMI for english dataset and **distant** bigrams

pair	PMI
(floated, dried)	8.692331
(dried, floated)	8.692331
(dried, dried)	8.303288
(floated, floated)	8.192257
(germinated, dried)	8.165785
(dried, germinated)	8.165785
(heath, heath)	8.060838
(clover, clover)	8.026448
(germinated, floated)	7.969865
(floated, germinated)	7.969865
(eastern, Pacific)	7.886809
(Pacific, eastern)	7.886809
(vibracula, avicularia)	7.854387
(avicularia, vibracula)	7.854387
(days, dried)	7.817861
(dried, days)	7.817861
(metamorphosed, metamorphosed)	7.817861
(cave, cave)	7.817861
(heads, heads)	7.805889
(THE, THE)	7.777219

2 Word classes

For the implementation of the Brown clustering algorithm, we follow the approach presented on slides 133–136 of the lecture. We constructed tables for mutual information, subtractions, and losses. Classes were merged based on minimal information loss and unigram and bigram counts were updated accordingly.

The full history of merges, along with the resulting 15 classes, is provided in the attached text files. For English dataset in files `history_EN.txt` and `15_classes_EN.txt`. Similarly for Czech dataset in files `history_CZ.txt` and `15_classes_CZ.txt`.

3 Tag classes

We attached full history of merges in file `history_tagsEN.txt` for English dataset. We decided to work with first 70000 tags for Czech data as the algorithm was time consuming. We attached history of merges in file `history_tagsCZ.txt`.

From history for English tags clustering we can see, that `comma`, `colon`, `dot` and `parantesis` are grouped together (steps 30, 21 and 15), suggesting a shared syntactic function in structuring sentences.

Another interesting group is CD + JJ (merge 24). This suggests that numbers behave like adjectives, which makes sense (e.g., three apples → ”three” describes ”apples”).

As the algorithm progresses, we observe that the WP (Wh-pronoun) class merges with the quotation mark (”) at step 29. One might expect quotation marks to merge with punctuation marks such as comma, colon, dot or parentheses instead. However, this grouping makes sense because Wh-pronouns and quotation marks often appear together in text. For example, in sentences like She asked, ”What is the answer?” or He said, ”Who told you that?”, the Wh-pronoun frequently follows an opening quotation mark, leading the algorithm to identify a strong association between them.