Comments and answers - Assignment 2

Project Text Analysis

Leon Wetzel, Teun Buijse and Roman Terpstra

Exercise 1

1.1c.

The ranking is different because all the top bigrams have the same score this is because their occurences together are the same. The chi square test eliminates more options because it takes into account how far observed values are from expected values, bacause of this the chi square tends to measure independance better than dependence and doenst the test account for how often the words do not appear together.

1.2.

When the range of both the rows and the columns in the table is the same. For example when having 2 rows and 2 columns in a 2x2 table in the case of bigrams when filling in the formula the rewritten formula can be found.

Exercise 2

2.1

Peter really liked the movies and warm pop-corn . He would never bring Mira with him, though.

Penn: NNP RB VBD DT NNS CC JJ NN . PRP MD RB VB NNP IN PRP , RP .

Brown: NP RB VBD AT NNS CC JJ NN . PPS MD RB VB NP IN PPO , RP .

Universal: NOUN ADV VERB DET NOUN CONJ ADJ NOUN . PRON VERB ADV VERB NOUN PRT PRON . ADV .

2.2

```
C:\Users\leonw\Documents\PTA\venv\Scripts\python.exe C:/Users/leonw/Documents/PTA/week2
        /exercise2.py
ASSIGNMENT 2 - PART 2
=============================================

2A. Amount of words and sentences...
- words: 57169
- sentences: 3886

2B. 50th and 75th words and tags...
grim: JJ (adjective)
from: IN (preposition)
```

```
2C. Amount of different POS tags...
169 different POS tags

2D. Top 15 words...
3326 .
2805 ,
2573 the
1284 to
1215 and
1136 a
903 of
820 was
740 ``
738 ''
670 he
664 ?
658 in
583 I
529 his

2E. Top 15 POS tags...
6461: NN (noun, singular, common)
4692: IN (preposition)
4322: . (sentence terminator)
4321: AT (article)
2805: , (comma)
2645: VBD (verb, past tense)
2459: RB (adverb)
2109: JJ (adjective)
2026: VB (verb, base: uninflected present, imperative or infinitive)
1767: PPS (pronoun, personal, nominative, 3rd person singular)
1737: NP (noun, singular, proper)
1692: CC (conjunction, coordinating)
1435: NNS (noun, plural, common)
1207: PPO (pronoun, personal, accusative)
1161: VBN (verb, past participle)

2F. Most frequent POS tag in the 20th and the 40th sentence...
1: PPSS+BER (pronoun, personal, nominative, not 3rd person singular + verb 'to be',
        present tense, 2nd person singular or all persons plural)
1: PPSS (pronoun, personal, nominative, not 3rd person singular)
Note: both sentences only contain unique tags and therefore every tag has a frequency of
        1.
The returned values are returned based on alphabetical order.

2G. Most frequent adverb...
Note: the set of adverbs is based on https://en.wikipedia.org/wiki/Brown_Corpus
up (193 occurences)

2H. Most frequent adjective...
Note: the set of adjectives is based on https://en.wikipedia.org/wiki/Brown_Corpus
old (58 occurences)

2I. POS tags for 'so'...
QL (qualifier, pre)
CS (conjunction, subordinating)
RB (adverb)
```

```
2J. Most frequent POS tag for 'so'...
QL - qualifier, pre (43 occurences)

2K. Example sentences for 'so'...
QL (qualifier, pre): We've always been so close '' .
RB (adverb): I'll bet he wouldn't be pleased if a rumdum like me were to ask his daughter
        for a date -- I mean , after I'm out of the hospital , a month or so from now ''
        .
CS (conjunction, subordinating): I put in new batteries so as to be certain I'd have
        plenty of power and on my way out walked over to the regular parking stalls and
        stood looking at them thoughtfully .

2L. Most likely POS tags preceding and following 'so'...
Most likely POS tag preceding 'so': , (comma)
Most likely POS tag following 'so': JJ (adjective)

Process finished with exit code 0
```

2.4

```
Top 5 significant:
[('FW-AT-TL', 'FW-NNS-TL'), ('FW-NNS-TL', 'FW-JJ-TL'), ('FW-PPL', 'FW-VB'), ('FW-VB', 'FW-
NP'), ('FW-VBZ', 'FW-RB')]
Top 5 raw:
[(('AT', 'NN'), 2692), (('IN', 'AT'), 1901), (('NN', 'IN'), 1379), (('NN', '.'), 1205),
(('JJ', 'NN'), 876)]
```

- Yes they look interesting, they're all foreign words which you would not expect.
- Yes they are quite different.
- This could for example be used in automatic text recognition. Foreign words are used more often in certain types of text, so together with other markers this could make automatic text recognition possible.