Basic NLP tools and concepts:Concordances, Vocabulary diversity, POS tagging, Tokenization

Dekel Viner

September 2016

partner: Ivo Steegstra

1 PART A: KWIK Concordances

1.1

1. Sense and Sensibility: 35
2. Wall Street Journal corpus:8

1.2

1. Sense and Sensibility:43
2. Wall Street Journal:20

1.3

Table 1: Part A: question 3

Meaning	Sense And Sensibility	Wall Street journal
Name	a more fascinating name ,**call** it hope	They **call** it " photographic " .
Phone	Why did you **call** , Mr.Willoughby ?	a telephone **call** to his office
Consider	do you **call** Colonel Brandon infirm	we would be hard put *-2 to **call** *T*-1 a game
Shout/cry	constrained to **call** for assistance	
ask to come	had sense enough to **call** one of the maids	
visit	and then they ' d be sure and **call** here	
decide on	determined not to **call** in Berkeley Street	action has been **called** for *-3 by an Argentine

1.4

1. for the definition of decide on : call in [(a) NP]. uses preposition.
2. for the definition of phone call: (specifier) called (complement)

2 Part B: Vocabulary diversity

2.1

I would think sense and sensibility would be the most lexically diverse as it is written by a novelist who generally enjoy using a broad vocabulary for higher descriptive purposes.

2.2

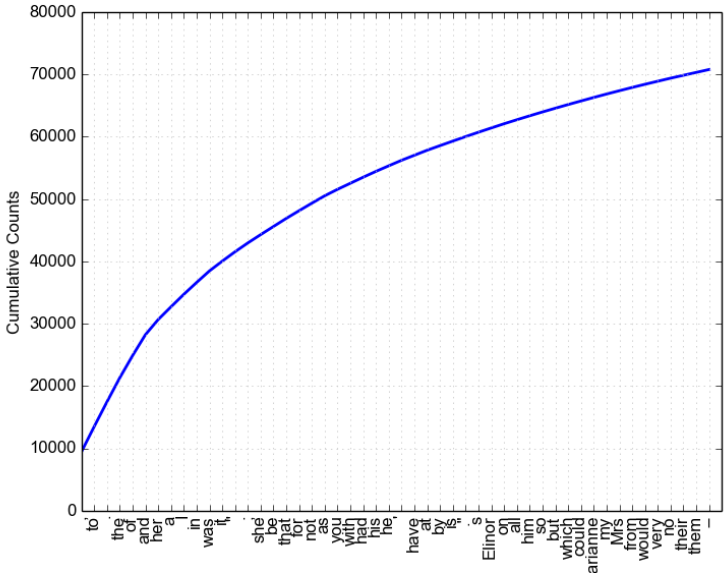
sense and sensibility: love
Chat corpus: hey
Wall Street Journal corpus: Economy

2.3

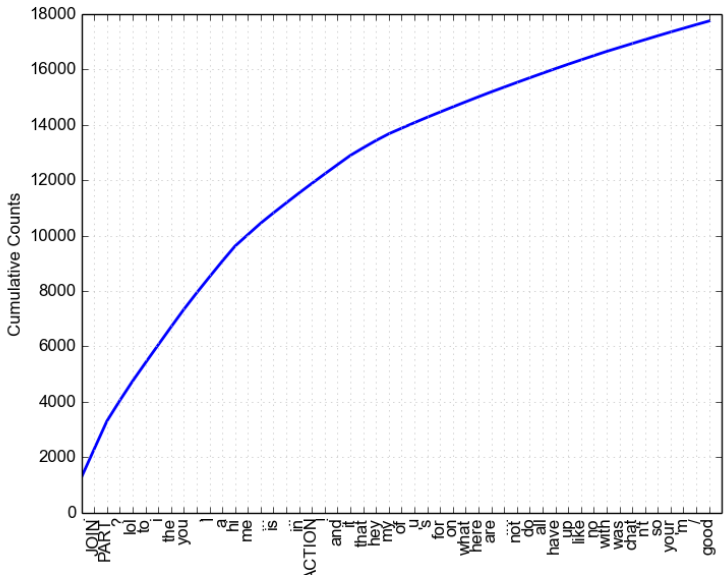
Table 2: Lexical Diversity

Corpus	Tokens	Type	Lexical Diversity
S&S	141576	6833	0.04826383003
Chat	45010	6066	0.1347700511
Wsj	100676	12408	0.1232468513

2.4



(a) SS top 50 word types



(b) Chat top 50 word types

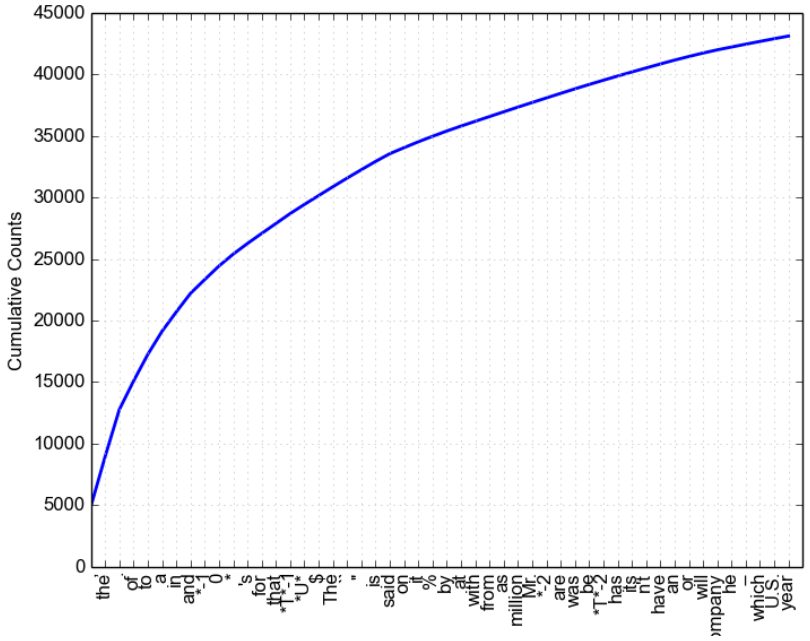


Figure 1: Wsj top 50 word types

2.5

Sense and sensibility: "Elinor" and "arianne" are highly typical of the text as they are names of the main characters likely.
Chat corpus: "hey" is highly typical as I imagined as it is an informal way of saying hello, and highly common starter of a chat conversation. Further more "lol", "hi", and "chat" seem to be also highly typical.
wall Street Journal corpus: company, the wall street journal is a business and economics focused journal there for it is no surprise that the word "company" would appear frequently. U.S also appears which would be typical of an United States based business journal.

3 Part C: Tokenization

The text:

Oh man, it's definitely so different. I also get sinus headaches relatively often (woo, allergies), and on top of that I get migraines maybe once a month (woo, menstruation). I can power through the sinus stuff. It sucks, but it's not debilitating. Migraines, though...I pretty much can't leave a dark, quiet space. I just lay there with tears streaming down my face, holding back sobs because that movement causes excruciating pain. And because it's a symptom of my period, I'm usually also experiencing cramps in my uterus and lower back pain as the fucking awful frosting on the cake of a living hell.

3.1

word Tokenizer: ['Oh', 'man', ',', 'it', "'s", 'definitely', 'so', 'different', '.', 'I', 'also', 'get', 'sinus', 'headaches', 'relatively', 'often', '(', 'woo', ',', 'allergies', ')', ',', 'and', 'on', 'top', 'of', 'that', 'I', 'get', 'migraines', 'maybe', 'once', 'a', 'month', '(', 'woo', ',', 'menstruation', ')', '.', 'I', 'can', 'power', 'through', 'the', 'sinus', 'stuff', '.', 'It', 'sucks', ',', 'but', 'it', "'s", 'not', 'debilitating', '.', 'Migraines', ',', 'though', '...', 'I', 'pretty', 'much', 'ca', "n't", 'leave', 'a', 'dark', ',', 'quiet', 'space', '.', 'I', 'just', 'lay', 'there', 'with', 'tears', 'streaming', 'down', 'my', 'face', ',', 'holding', 'back', 'sobs', 'because', 'that', 'movement', 'causes', 'excruciating', 'pain', '.', 'And', 'because', 'it', "'s", 'a', 'symptom', 'of', 'my', 'period', ',', 'I', "'m", 'usually', 'also', 'experiencing', 'cramps', 'in', 'my', 'uterus', 'and', 'lower', 'back', 'pain', 'as', 'the', 'fucking', 'awful', 'frosting', 'on', 'the', 'cake', 'of', 'a', 'living', 'hell', '.']

white space Tokenizer: ['Oh', 'man,', "it's", 'definitely', 'so', 'different.', 'I', 'also', 'get', 'sinus', 'headaches', 'relatively', 'often', '(woo,', 'allergies),', 'and', 'on', 'top', 'of', 'that', 'I', 'get', 'migraines', 'maybe', 'once', 'a', 'month', '(woo,', 'menstruation).', 'I', 'can', 'power', 'through', 'the', 'sinus', 'stuff.', 'It', 'sucks,', 'but', "it's", 'not', 'debilitating.', 'Migraines,', 'though...I', 'pretty', 'much', "can't", 'leave', 'a', 'dark,', 'quiet', 'space.', 'I', 'just', 'lay', 'there', 'with', 'tears', 'streaming', 'down', 'my', 'face,', 'holding', 'back', 'sobs', 'because', 'that', 'movement', 'causes', 'excruciating', 'pain.', 'And', 'because', "it's", 'a', 'symptom', 'of', 'my', 'period,', "I'm", 'usually', 'also', 'experiencing', 'cramps', 'in', 'my', 'uterus', 'and', 'lower', 'back', 'pain', 'as', 'the', 'fucking', 'awful', 'frosting', 'on', 'the', 'cake', 'of', 'a', 'living', 'hell.']

3.2

The word toknizer tokenizes "(woo, allergies)" as " '(', 'woo', ',', 'allergies', ')' " while the white space tokenizer has " '(woo,', 'allergies),' ". So the word tokenizer seperates the words from the parenthesis while the the white space tokenizer does not. I believe the word tokenizer is more correct in this case, as the parenthesis are not a part of the word. The word tokenizer tokenizer "it's" as " 'it', "'s", " and the white space tokenizer just has "it's". Regarding which one is more correct I believe it depends on the purpose of the tokenisation, but both can be deemed correct.

4 Part D: Part-of-speech tagging

4.1

[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'), ('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN'), ('.', '.')]

One potential problem is that the two "refuse" cases are two different word types. And a similar case with "permit" and "permit" which in this context are two completely different words. The POS tagger recognizes the first refuse correctly as a verb because it follows the personal pronoun "they", so it must be a verb. The second "refuse" follows a determiner "the" so it must be a noun. The first "permit" is categorized correctly as a verb as it follows 'to', while the second "permit" is categorized correctly as a noun, the POS tagger likely recognizes that it follows another noun and therefore does not consider it a verb.

4.2

Generated by the POS tagger:

[('Time', 'NNP'), ('flies', 'NNS'), ('like', 'IN'), ('an', 'DT'), ('arrow', 'NN')] (As generated by the POS tagger)

Generated by us:

[('Time', 'VBP'), ('flies', 'NNS'), ('like', 'IN'), ('an', 'DT'), ('arrow', 'NN')]

First sentence: Time flies are a specific kind of flies, that have properties similar to an arrow. Time flies that are like an arrow. Comparable sentence: House cats like a tiger. The sentence seems incomplete.

Second sentence: Time is now a verb, you 'time' an activity that the flies do, in a similar way that you 'time' something an arrow does (like fly from one point to another).

4.3

Sense and Sensibility:
'NN' = 16680; Example: "hand"
'IN'= 15511; Example: "by"
'PRP'= 10893; Example: "she"
'DT'= 9560; Example: "the"
'RB', 8506; Example: "very"

Wall Street Journal:

'NN', 14666; Example: "agenda"
'NNP', 10457; Example: "U.S."
'IN', 10055; Example: "with"
'JJ', 8747; Example: "spectacular"
'DT', 8117; Example: "some"

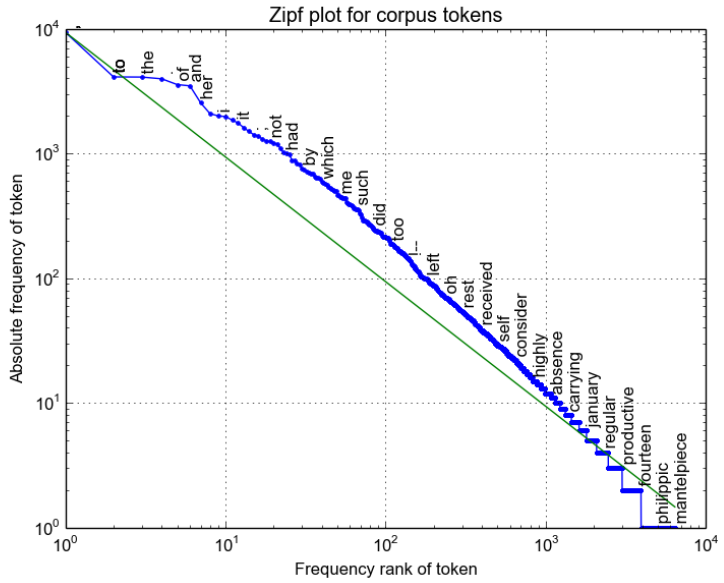
4.4

They are not the same Sense and Sensibility has more personal pronouns, this seems to make sense as it is a novel, and the writer uses a lot of pronouns to refer to the characters in the book. Another difference is regarding subordinate conjunctions (IN), Sense and Sensibility seems to have a lot more of them. The role of subordinate conjunctions is to introduce a subordinating clause (aka dependent clause). Subordinating clauses are not typical of journal articles where most of the writing is focused mainly on stating facts while being concise and to the point, however they do add to creative writing, story telling and to the flow of the writing and therefore makes them more prominent in novels. Adverbs(RB) appear to be more prominent in Sense and Sensibility and of no surprise. Adverb are used in novels for descriptive purposes.

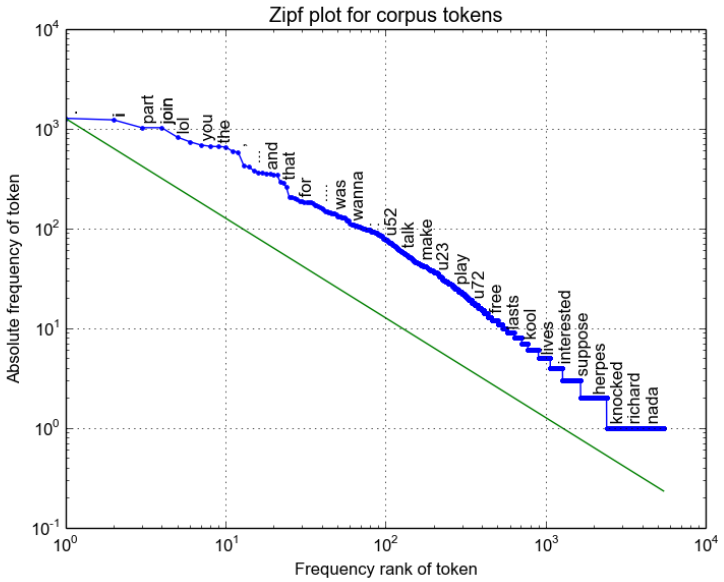
proper nouns appear to be very frequent in the Wall Street Journal, that is of no surprise names of companies, products, countries and people would all appear as proper nouns. It hard to imagine that proper nouns would not also be common in novels, it is possible that their lower frequency in Sense and Sensibility is due to the writer style of writing or is a mere coincidence. A similar case with adjectives, you could imagine that a different book would have a higher rank for adjectives.

5 Part E: Zipfian distributions

5.1



(a) SS log Zipf



(b) Chat log Zipf

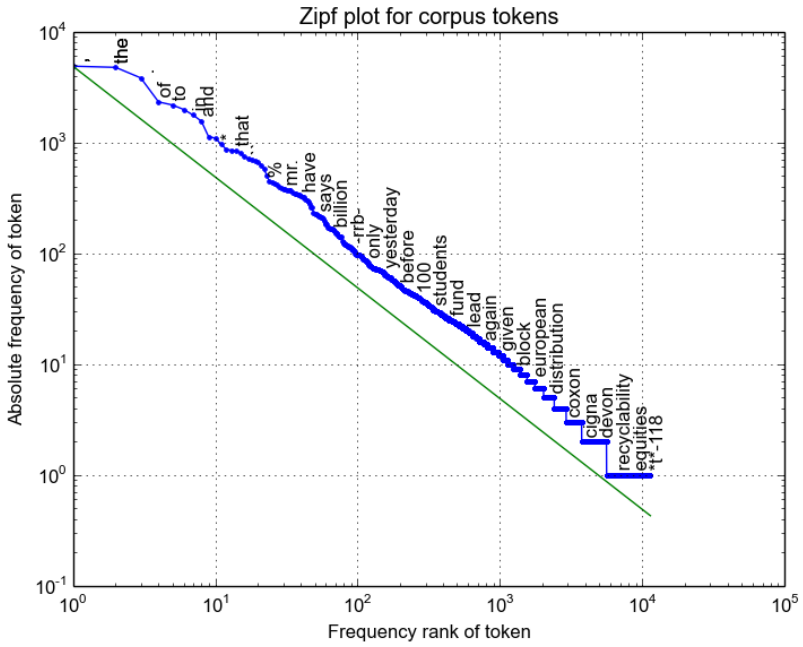


Figure 2: Wsj log Zipf

5.2

Table 3: Text2 zipf prediction vs actual

Rank	Word type	Frequency	Predicted Frequency
1	to	4116	4166
2	the	4105	2058
3	of	3572	1372
4	and	3491	1029
5	her	2551	823.2

The zipf relationship predicts that the second most frequent word type would be half as frequent as the first, and the third half as frequent as the second, and every rank after that is half as frequent as the rank before.
The relationship does not accurately hold for most part which can be also seen when considering the curves. The curves however also contain symbols that are not word types. Still, it appears that the actual frequencies are higher than the predicted frequencies. With the wsj(text7) the predicted frequency is the closest to the actual frequency

Table 4: Text5 zipf prediction vs actual

Rank	Word type	Frequency	Predicted Frequency
1	i	1224	1224
2	part	1022	612
3	join	1021	408
4	lol	822	306
5	you	686	244.8

Table 5: Text7 zipf prediction vs actual

Rank	Word type	Frequency	Predicted Frequency
1	the	4764	4764
2	of	2325	2382
3	to	2182	1588
4	a	1988	1191
5	in	1769	952.8