Central Bank Communications: Information extraction and Semantic Analysis.

Giuseppe Bruno[1]

[1]Economics and statistics Directorate
Bank of Italy

Harnessing Big Data & Machine Learning Technology for Central Banks, Rome. March 26-27

Outline

Outline

Availability of textual data

Extracting information from textual data

- Central Institutions express their position through documents as well as quantitative figures.
- The web provides an enormous warehouse of information. Around 4/5 of this info is of textual nature.
- Harnessing textual information requires a theoretical approach. We adopted the *bag of words* assumption.

Availability of textual data

Extracting information from textual data

- Central Institutions express their position through documents as well as quantitative figures.
- The web provides an enormous warehouse of information. Around $4/5$ of this info is of textual nature.
- Harnessing textual information requires a theoretical approach. We adopted the *bag of words* assumption.

Availability of textual data

Extracting information from textual data

- Central Institutions express their position through documents as well as quantitative figures.
- The web provides an enormous warehouse of information. Around $4/5$ of this info is of textual nature.
- Harnessing textual information requires a theoretical approach. We adopted the *bag of words* assumption.

Building a Corpus of Text documents.

Text format is the starting point for any kind of textual and semantic analysis.
A set of preprocessing tasks are usually required for building a useful corpus:

a) lowercase conversion and white space removal;
b) stopwords and numbers elimination;
c) stemming or lemmatization;
d) special characters conversions or filtering;

The Bag of words model

Consider the following two sentences:

s1: I rischi per la stabilita finanziaria sono attenuati.
s2: La BCE ha interrotto la spirale negativa tra aumento dei rischi sovrani e difficolta del sistema bancario.

A set of 23 words for a list a 22 distinct words

Our vocabulary is a matrix with 2 rows and 22 columns: the word and the its number of occurrence

i, rischi, per, la, stabilita, finanziaria, sono, attenuati, BCE, ha, interrotto, spirale, negativa, tra, aumento, dei, sovrani, e, difficolta, del, sistema, bancario

s1: 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
s2: 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

Each sentence is represented by a vector: s1 or s2

The representation, referred to as the bag-of-words representation, is not faithful, as it ignores the respective order of appearance of the words. In addition, often, stop words (such as articles and prepositions) are ignored.

Bag of Words model: Accuracy and Recall

Bag of Words is unfaithful but:
it often allows applications with good accuracy and recall in classification.

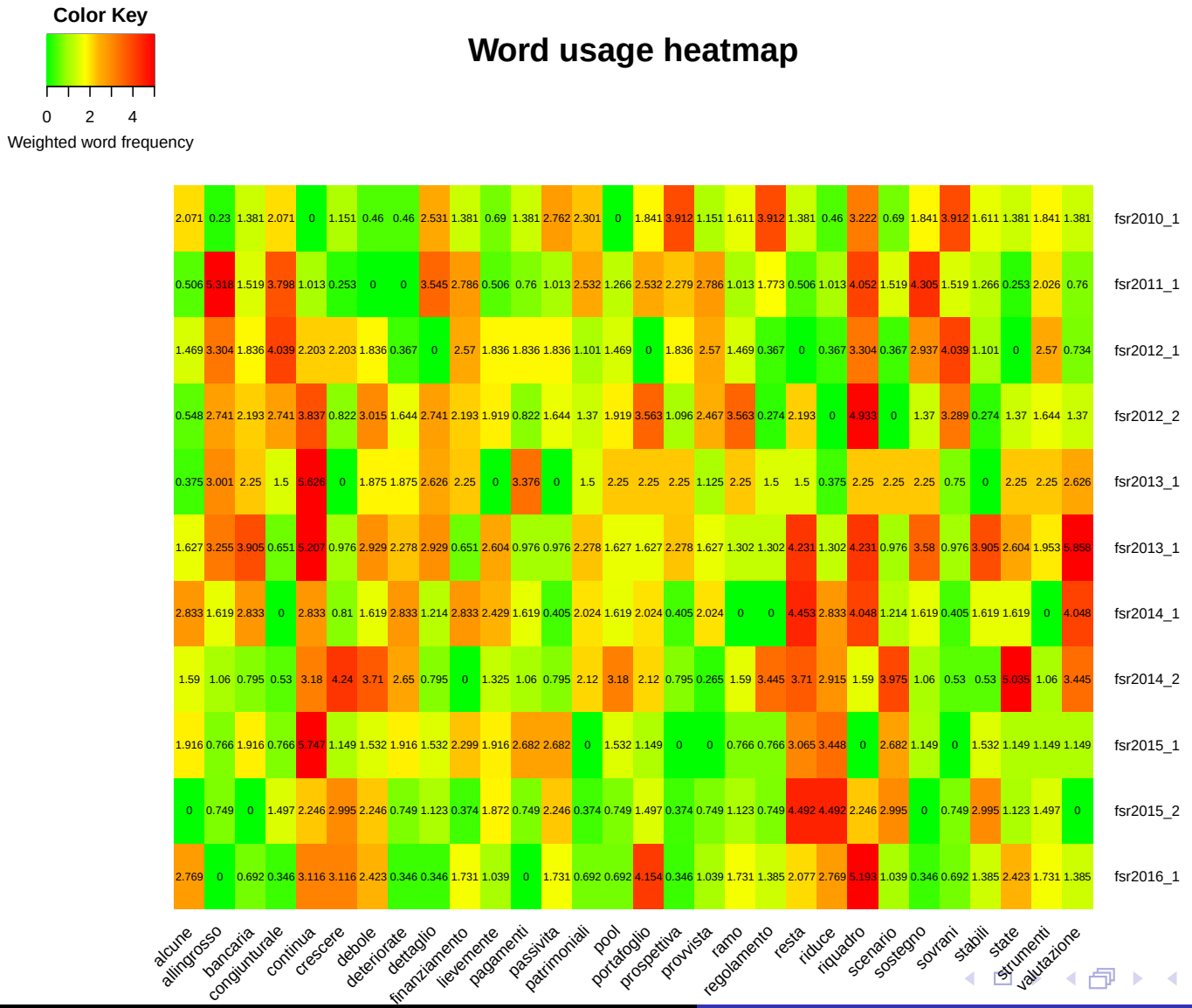
$$accuracy = \frac{true\_positive}{true\_positive + false\_positive}$$

$$recall = \frac{true\_positive}{true\_positive + false\_negative}$$

Statistics for the Financial stability report

issue	#sentence	#word per sentence	sd #word	#char per sentence	#char per word
2010_1	518	31.30	14.69	182.41	5.83
2011_1	428	32.40	15.29	190.00	5.86
2012_1	295	32.97	16.27	191.99	5.82
2012_2	364	33.18	16.06	192.01	5.78
2013_1	288	32.21	15.56	187.26	5.81
2013_2	317	31.85	15.46	185.60	5.83
2014_1	271	31.52	15.10	181.26	5.75
2014_2	379	34.21	16.64	195.40	5.71
2015_1	266	34.32	14.98	195.94	5.71
2015_2	267	32.21	14.92	183.88	5.71
2016_1	297	32.87	14.94	187.57	5.71

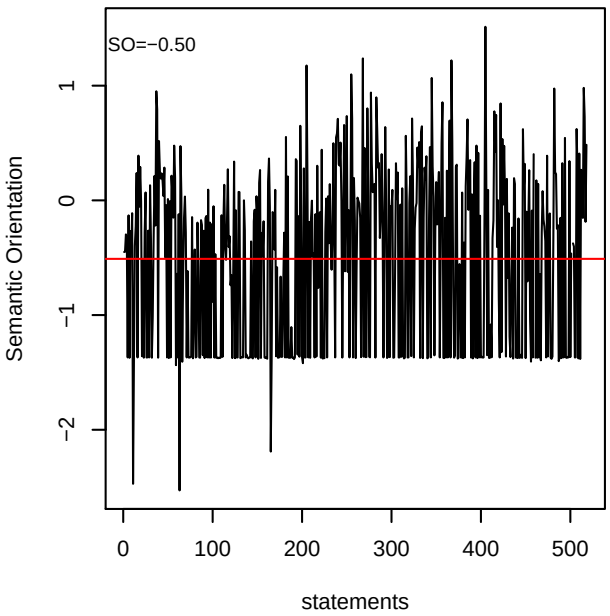
Outline


Color Key
0
2
4
Weighted word frequency
Word usage heatmap
fsr2010_1
fsr2011_1
fsr2012_1
fsr2012_2
fsr2013_1
fsr2013_1
fsr2014_1
fsr2014_2
fsr2015_1
fsr2015_2
fsr2016_1
alcune
allingrosso
bancaria
congiunturale
continua
crescere
debole
deteriorate
dettaglio
finanziamento
lievemente
pagamenti
passivita
patrimoniali
pool
portafoglio
prospettiva
provvista
ramo
regolamento
resta
riduce
riquadro
scenario
sostegno
sovrani
stabili
state
strumenti
valutazione

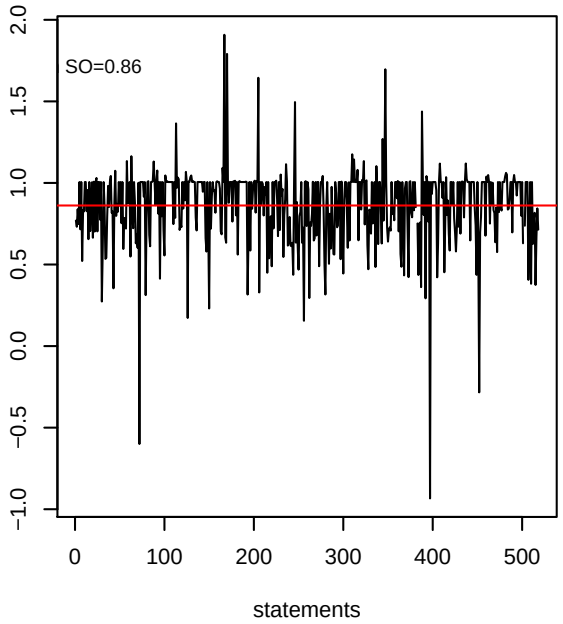
Semantic Orientation in 2010

Semantic Orientation in 2010_1

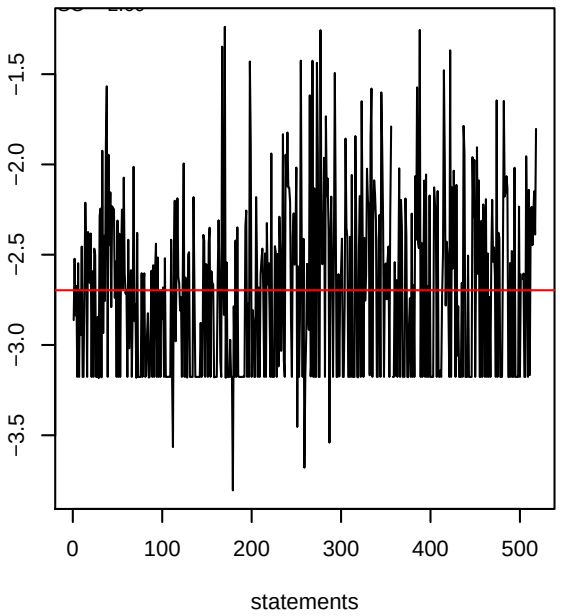
Antinomy stabilità/instabilità



Antinomy espansione/crisi



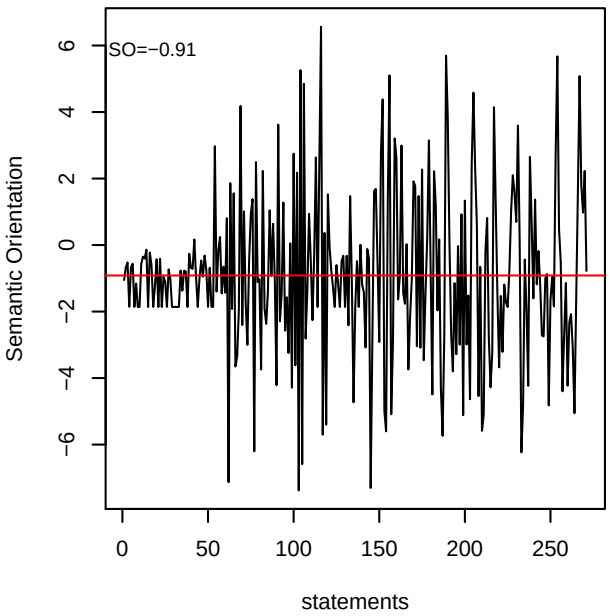
Antinomy solidità/vulnerabilità



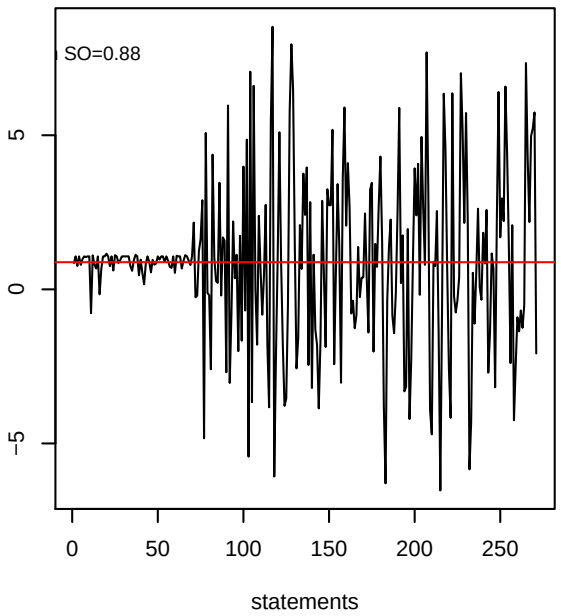
Semantic Orientation in 2014

Semantic Orientation in 2014_1

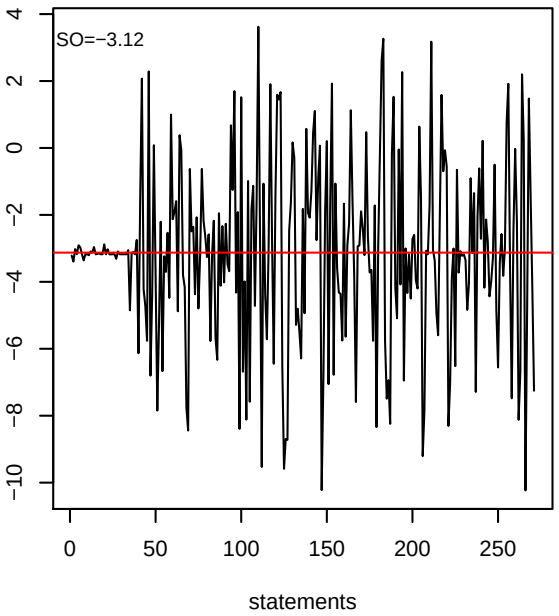
Antinomy stabilità/instabilità



Antinomy espansione/crisi



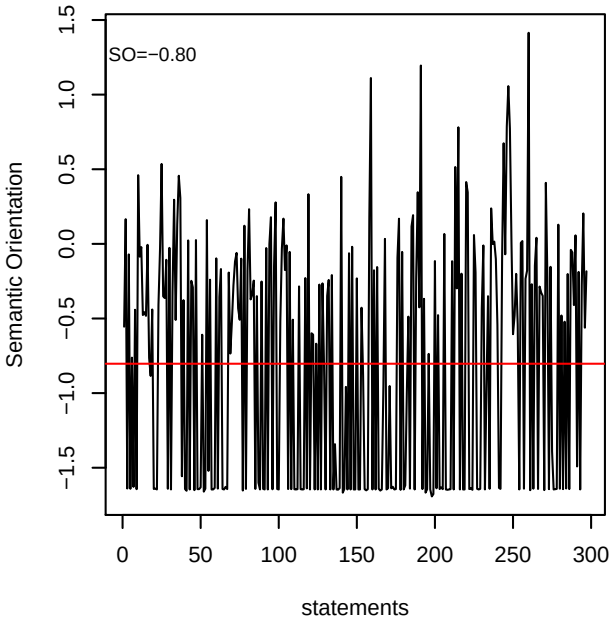
Antinomy solidità/vulnerabilità



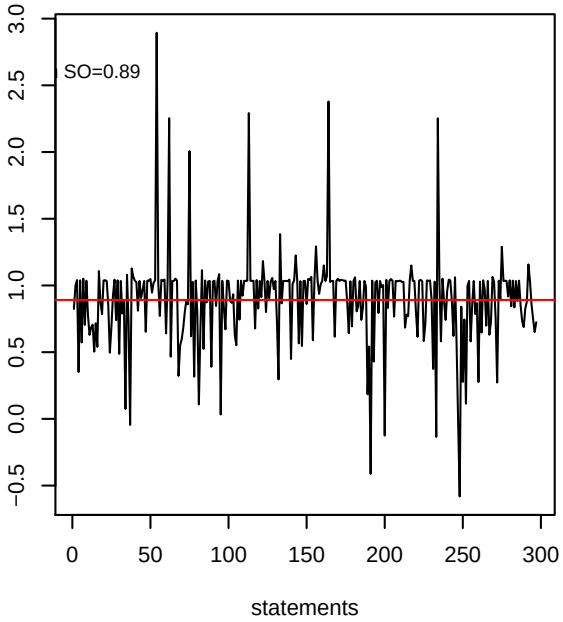
Semantic Orientation in 2016

Semantic Orientation in 2016_1

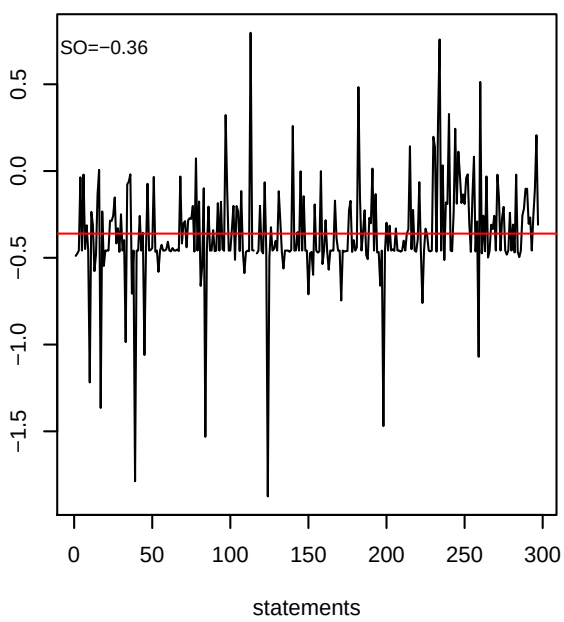
Antinomy stabilità/instabilità



Antinomy espansione/crisi



Antinomy solidità/vulnerabilità



Outline

Measuring the sentiment of a sentence/document.

Once a text document has been suitably split into a set of sentences, it is possible to apply sentiment extraction algorithms. These algorithms take into account the following:

- polarized words in the statement. These can be adjective or adverbs ;
- adjective amplifiers (comparative & superlative);
- modifiers like negations.

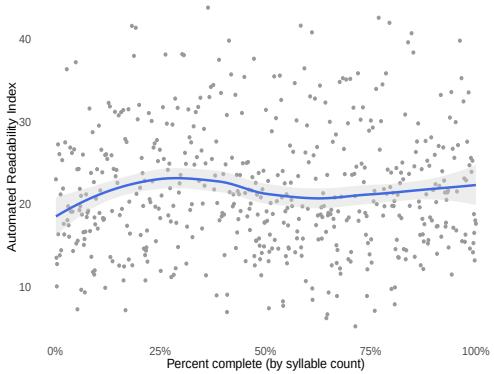
Outline

Readability assessment provides a measure of the effort required by a reader to understand a text.
Readability is a shallow structure of the text and can be extracted by simply counting words and characters.

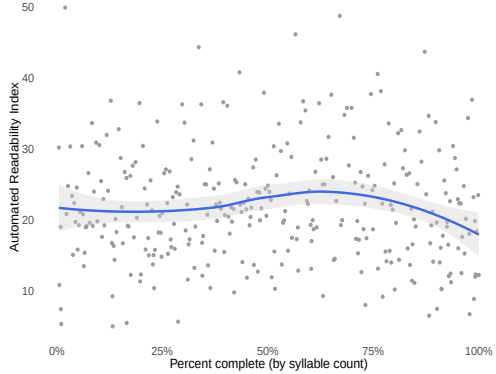
There are at least six different definitions of readability. We have adopted the Automated Readability Index *ARI* which is aimed at the English language

$$ARI = 4.71 \cdot \left(\frac{N_{char}}{N_{words}} \right) + .5 \cdot \left(\frac{N_{words}}{N_{sentences}} \right) - 21.43$$

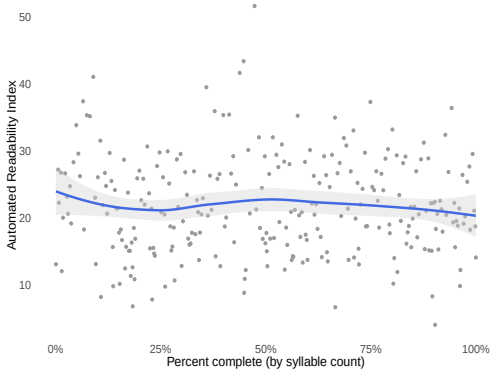
This index, available in the R **qdap** package, rewards shorter words and sentences.



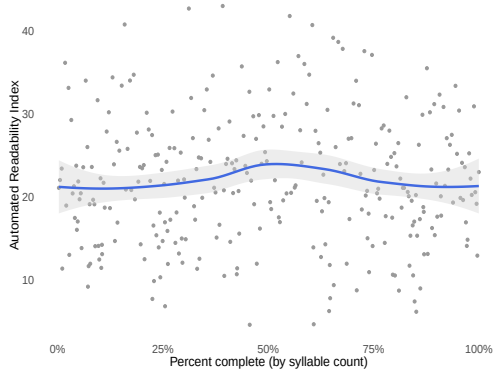
Readability FSR 2010



Readability FSR 2013-2



Readability FSR 2015-2



Readability FSR 2016-1

Outline

The Formality measure.

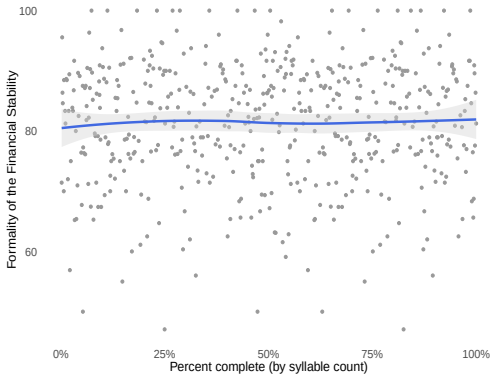
Formality of a statement/text is defined as the amount of expression that is immutable irrespective to changes of context. Examples come from the consideration of spatial-temporal context.
"Today Tom is there" vs "The 5th of October 2016, Tom is at the Bank of Italy"

The Formality measure.

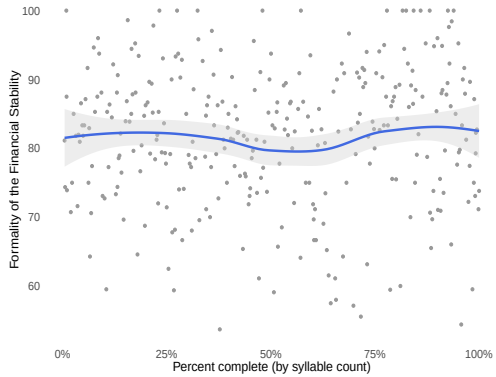
Following Heylighen and Dewaele (2002) the formality is computed as:

$$F = 50 \cdot \left(\frac{n_f - n_c}{N} + 1 \right) \tag{1}$$

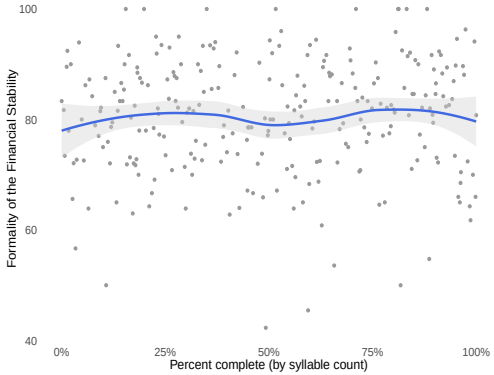
where n_f is the total number of nouns, adjectives, prepositions and articles, and n_c is the total number of pronouns, adverbs, verbs and interjections. The normalizing constant $N = \sum (f + c + \textit{conjunctions})$



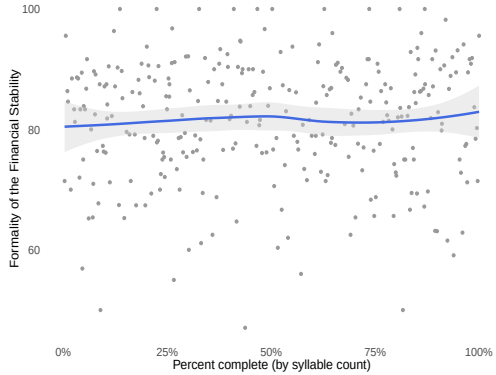
Formality of FSR 2010



Formality of FSR 2013-2



Formality of FSR 2015-2



Formality of FS 2016-1

Outline

Semantic Orientation from PMI

We can infer semantic orientation from semantic association. The semantic orientation of a given word/sentence is calculated from the strength of its association with a set of positive words, minus the strength of its association with a set of negative words

$$SO(sent) = \sum_{pos\_wd} (A(sent, pos\_wd)) - \sum_{neg\_wd} (A(sent, neg\_wd))$$

Semantic Orientation from PMI

Given two events x and y, we consider the ratio between the joint probability of co-occurrence and the product of the probabilities of the two events:

$$PMI(x;y) \equiv \log \frac{p(x,y)}{p(x) \cdot p(y)} \tag{2}$$

PMI measures the degree of statistical independence between x and y.

Semantic Orientation from PMI

The semantic orientation of a sentence $sent_i$ is then evaluated as

$$SO(sent_i) \equiv PMI(sent_i, ' wonderful') - PMI(sent_i, ' awful')$$

Here we compare which adjective is more informative in explaining the sentence.

Semantic Orientation from PMI

The semantic orientation can be made more robust by employing and array of N antonyms:

$$SO(sent_i) \equiv \sum_{ant_j=1}^{N} PMI(sent_i, ant_j[positive]) - PMI(sent_i, ant_j[negative]$$

Here we try to shrink the estimation variance by taking advantage of different couple of polar words.

Outline

Concluding Remarks

- We have written some *R* functions for building a Corpus of Central Bank documents;
- we have measured the adherence of these documents to the Zipf's and Heaps' law;
- we have evaluated some general characteristics of these documents (readability and formality);
- we have made a first attempt in evaluating the sentiment and polarity orientation in the text.

For Further Reading


F. Heylighen and J. Dewaele.
Variation on the Contextuality of Language: an Empirical Measure.
Foundation of Science, 2002.


R. Senter and E.A. Smith.
Automated Readability Index.
Aerospace Medical Research Laboratory, 2010.


L. Egghe.
Untangling Herdan's Law and Heaps Law: Mathematical and Informetric Arguments.
Journal of the American society for Information Science and Technology, 2007.

Thank you for your attention.

Any questions?