Central Bank Communications: Information Extraction and Semantic Analysis.

Giuseppe Bruno, Bank of Italy

Abstract

Central Banks, among other tasks, provide a relevant amount of information for Institutions and market operators. Indeed central banks employ a multiplicity of communication channels to drive market expectations.
In this paper we present some methodologies aimed to quantify the information content of official communications and we present their application to the semi-annual publication of the Financial stability report. While these methodologies are quite developed for the English and other highly spoken languages in the world, they are still in their experimental phase for the Italian language.
Here the goal is twofold: on one hand we provide a transparent numerical framework to consider sub-unit of an official Central Bank report written in Italian. Moreover it is proposed an analytical tool to gauge the impact of an official document on the public.
In the context of reports released by the Bank of Italy, we show how this framework can be employed to numerically characterize and extract their information content.
We deem quite relevant a quantitative evaluation of the impact of these reports in increasing the central bank transparency with the goal of enhancing the effectiveness of its institutional action.

JEL classification: C83, E58, E61
Keywords: Text Mining, Semantic Analysis, Pointwise Mutual Information, Web search.

Outline

1) Motivation
2) Corpora of documents and their statistical features
3) Shallow and Syntactic features of documents (Readability and Formality)
4) Latent Semantic Analysis
5) Pointwise Mutual Information and Semantic Orientation (Sentiment on a given topic & Web Hit approximated PMI)

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. Here we adopted the bag of words assumption.

Issue	#sentences	#word per sentences	#sd word	#char per sentence	#char per word
2010_1	518	31.3	14.69	182.41	5.83
2011_1	428	32.4	15.29	190	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	185.6	5.83
2014_1	271	31.52	15.1	181.26	5.75
2014_2	379	34.21	16.64	195.4	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

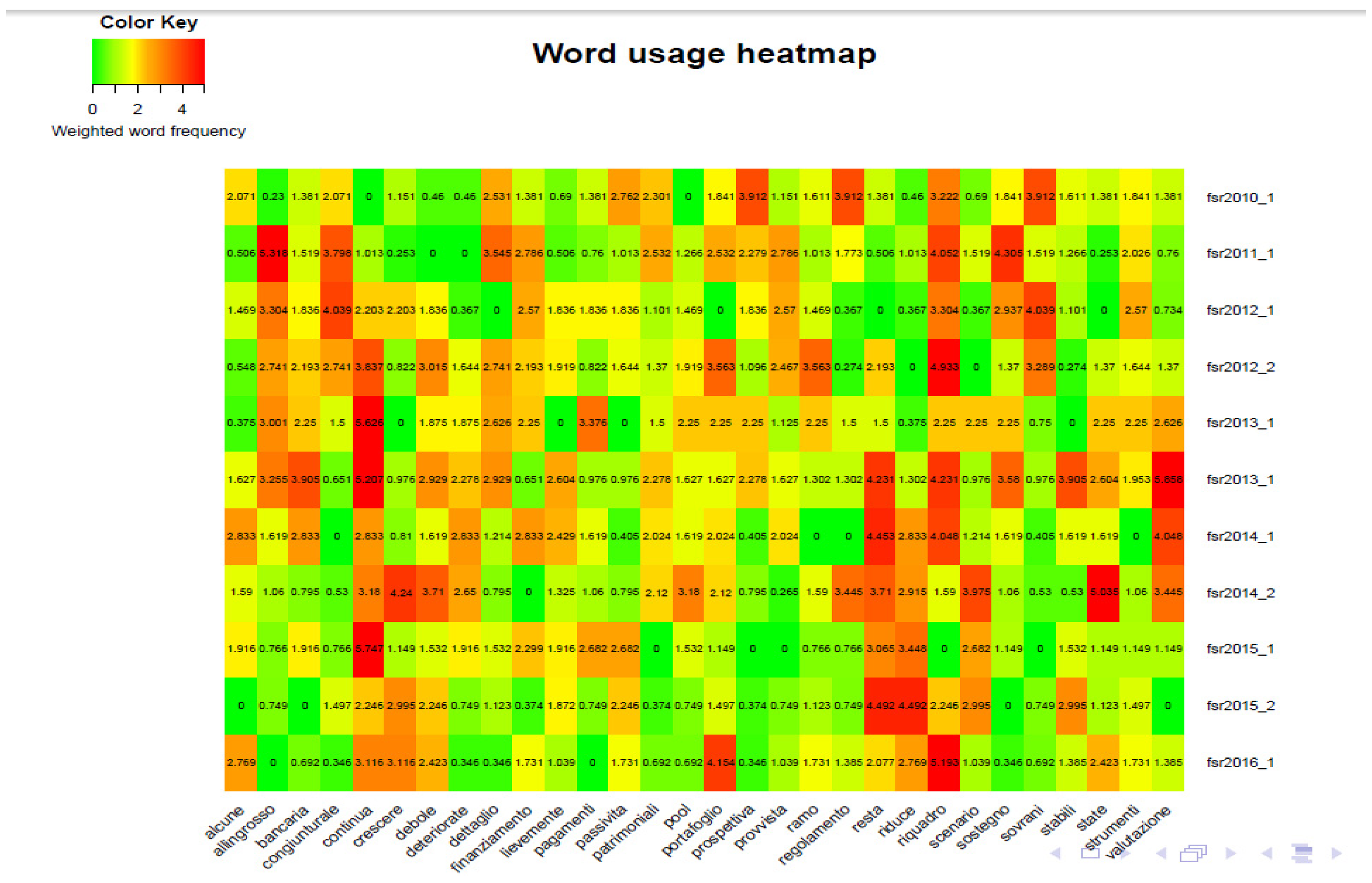
How to get a heatmap?

```
require('gplots')
heatmap.2(matfsr,
cellnote = matfsr,                     # same data set for cell labels
notecex=0.6,                           # size of note in the cell
main = "Word usage heatmap",           # heat map title
notecol="black",                       # change font color of cell labels to black
density.info="none",                   # turns off density plot inside color legend
key=TRUE, symbreaks=FALSE,
trace="none",                          # turns off trace lines inside the heat map
margins =c(5,16),                      # widens margins around plot
col=my_palette,                        # use on color palette defined earlier
breaks=col_breaks,                      # breaks in color changing
dendrogram="none",                     # only draw a row dendrogram
cexCol=.9,                             # specify row label font size
cexRow=.81,                            # specify row label font size
srtCol=45,                             # rotate the column labels
lmat=rbind( c(4, 3), c(2,1) ), lhei=c(0.18, .7 ), lwid=c(0.65,4),
key.xlab="Weighted word frequency",
Colv=FALSE,Rowv=FALSE)                 # turn off column clustering
```

Quite a few number of parameters to set!!

The Heatmap for a set of documents

A global diagnostic tool for a corpus of homogeneous documents is the heatmap.
A heatmap provides a picture showing the hottest word (more used) for different documents.

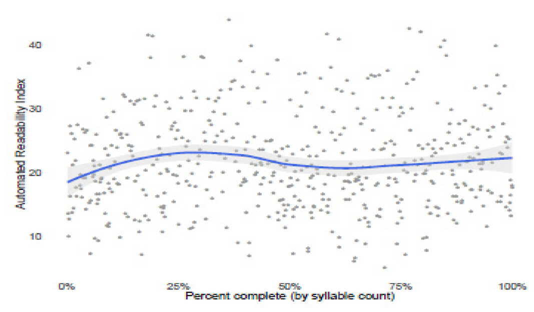


The Readability definition

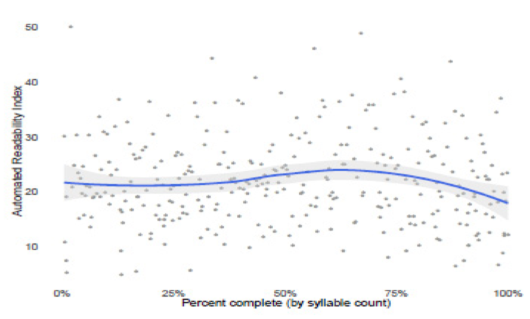
Readability assessment provides a measure of the effort required by a reader to understand a text.
Readability is a shallow feature of the text and can be extracted by simply counting words and characters.
There are at least six different definitions of readability. We have adopted the Automated Readability Index ARI which is aimed at the English language

$$ARI = 4.71 \cdot \left(\frac{N_{char}}{N_{words}}\right) + 0.5 \cdot \left(\frac{N_{words}}{N_{sentences}}\right) - 21.43$$

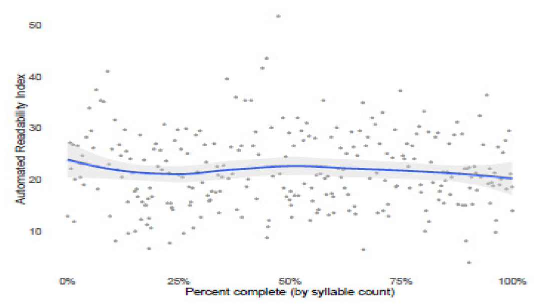
This index, available in the **qdap** package, rewards shorter words and sentences.



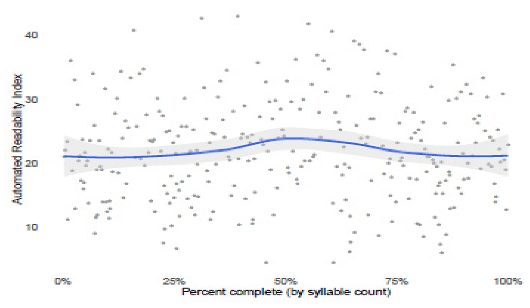
Readability FSR 2010



Readability FSR 2013-2



Readability FSR 2015-2



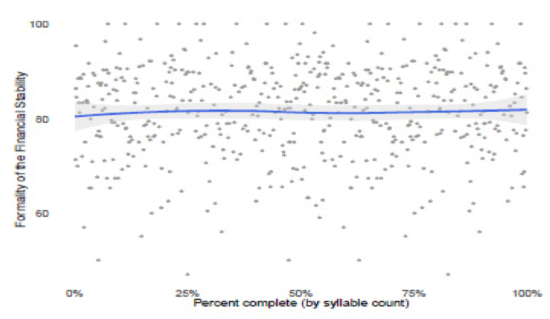
Readability FSR 2016-1

The Formality definition

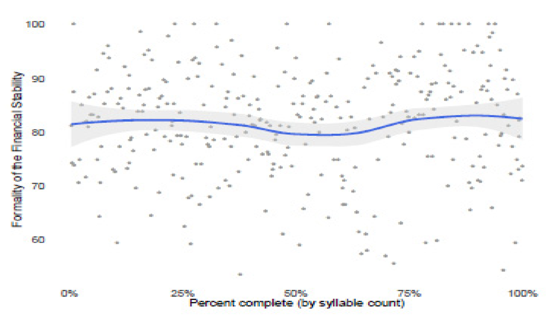
Formality of a statement/text is defined as the amount of expression that is immutable irrespective to changes of context.
Examples come from the consideration of spatial-temporal context.
"Today Tom is there" vs "The 5th of October 2016, Tom is at the Bank of Italy". Formality is computed according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

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

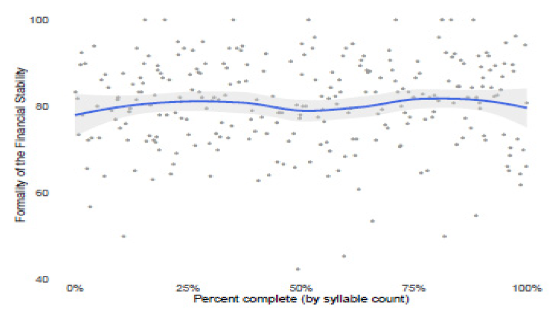
where: n_f is the total number of nouns, adjectives, prepositions and articles, and n_c is the total number of pronouns, adverbs, verbs and interjections. The normalizing constant is given by $N = \sum(f + c + conjunctions)$



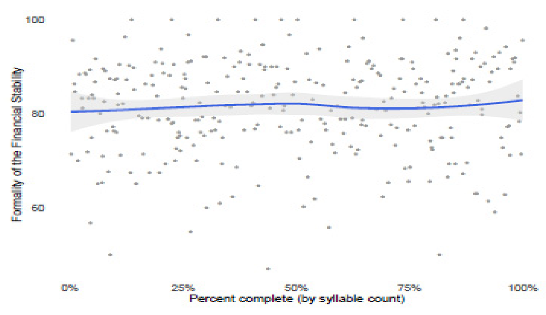
Formality of FSR 2010



Formality of FSR 2013-2



Formality of FSR 2015-2



Formality of FS 2016-1

Latent Semantic Analysis

After completing the task of building a corpus of documents, it is possible to start the semantic analysis.
Latent Semantic Analysis (LSA) is a methodology for extracting and representing the contextual-usage of words (co-occurrence) for determining the similarity of meaning of sentences by analysis of large text corpora.
The input for the LSA algorithm is a text document matrix:

$$TDM = \begin{array}{c} \\ word_1 \\ word_2 \\ \vdots \\ word_n \end{array} \begin{array}{c} \begin{array}{cccc} doc_1 & doc_2 & \cdots & doc_n \end{array} \\ \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{1,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{pmatrix} \end{array}$$

each $w_{i,j}$ is a weighted value of the number of occurrences of the word *i* in document j.

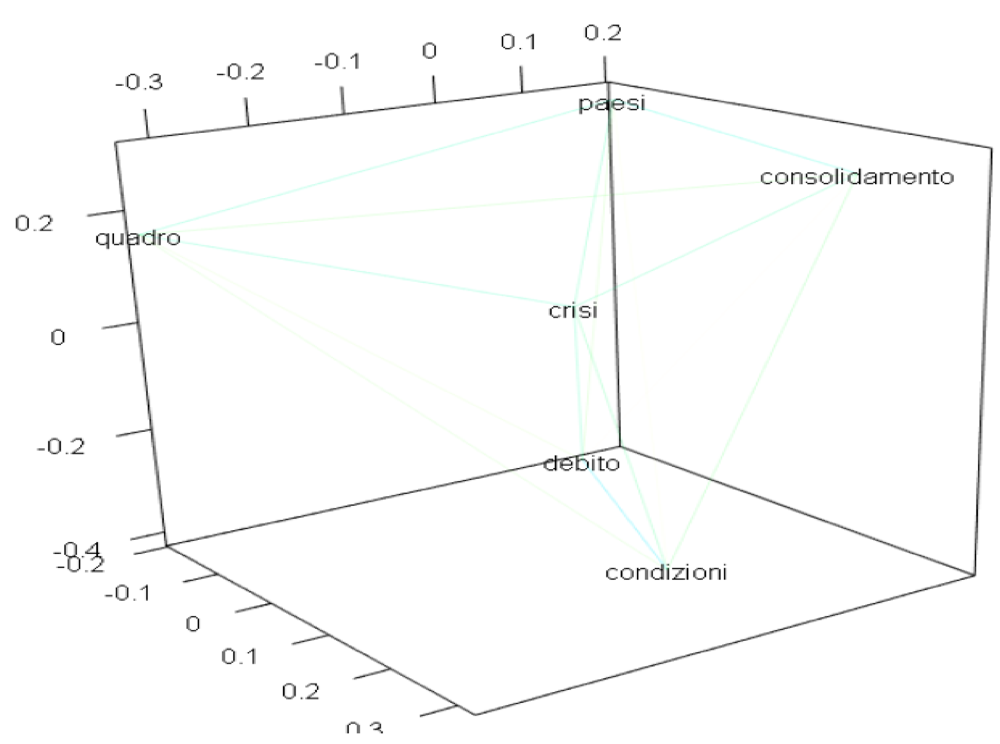
The TDM matrix is decomposed with the Singular Value Decomposition procedure:

$$TDM = U \cdot \Sigma \cdot V^t$$

Here the trick is that U and V are orthonormal matrices. Orthogonal basis implies the ability to decompose an effect into separate, non-interacting parts that simply add up to form the whole effect. This is a generalization of the Factor Analysis.

Latent Semantic Analysis applications

Words most highly similar with 'crisi'



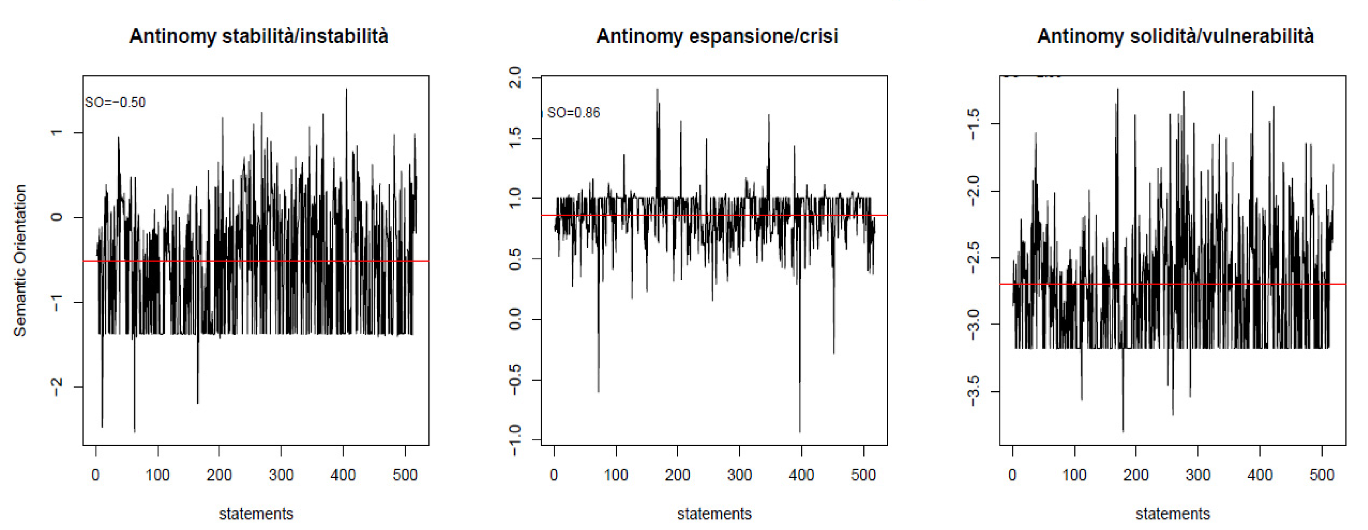
Semantic Orientation from PMI

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:

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

Each one of the sums is approximated as $\sum_{pos\_wd}(A(sent, pos\_wd)) \equiv PMI(sent; pos\_wd)$ and $PMI(x;y) \equiv log\frac{p(x,y)}{p(x) \cdot p(y)}$

Semantic Orientation in 2010_1



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.

D. Lucca and F. Trebbi.
Measuring Central Bank Communication: an Automated Approach with Applications to FOMC Statements.
NBER working paper, 2011.