SpeechGraphs v 1.0

1. Introduction

This document is a guide for the free SpeechGraphs software, a graph-theoretical analysis tool that uses text as input and graph features as output. SpeechGraphs can run on many different platforms such as Linux, Windows and OSX. This guide will be updated each time we release a new version of SpeechGraphs. You should receive a copy of this guide along with the software; alternatively, this guide can be directly downloaded from the project website (http://www.neuro.ufrn.br).

2. Dependencies

SpeechGraphs uses the following third-party libraries:

- commons-math-2.2.jar (Apache 2.0 License);
- collections-generic-4.01.jar (Apache 2.0 License);
- JUNG 2.0.1 (BSD License) - Java Universal Network/Graph Framework (http://jung.sourceforge.net/);
- JBLAS 1.2.4 (BSD License) - https://github.com/mikiobraun/jblas;
- JFreeChart 1.0.14 (LGPL V3.0) - https://sourceforge.net/projects/jfreechart/files/;
- JCommon 1.0.17 (LGPL V3.0) - https://sourceforge.net/projects/jfreechart/files/;
- Eclipse Jar-in-Jar Loader (EPL V2.0) – https://github.com/eclipse/eclipse.jdt.ui;
- Freehep VectorGraphics 2.1.1 (LGPL V2.1) - http://java.freehep.org/vectorgraphics/.
- GraphStream gs-core 1.3 and gs-algo 1.3 (LGPL V3.0) - http://graphstream-project.org/.

3. Getting Started

Here we describe how to set up the SpeechGraphs environment and get it running.

3.1 Requirements

• Java Virtual Machine (JVM) installed.

3.2 Get it running

After downloading SpeechGraphs, double-click the icon of the application.

4. Approaches

4.1. Visualization

4.1.1 Visualization File Format

The Visualization file format is a simple text file, as shown below:

... ah o sonho mais recente que eu lembro é eu em noronha mais minha rapaziada meus amigos todos até uns dois que neste sonho surfando bastante com uma prancha que nós …

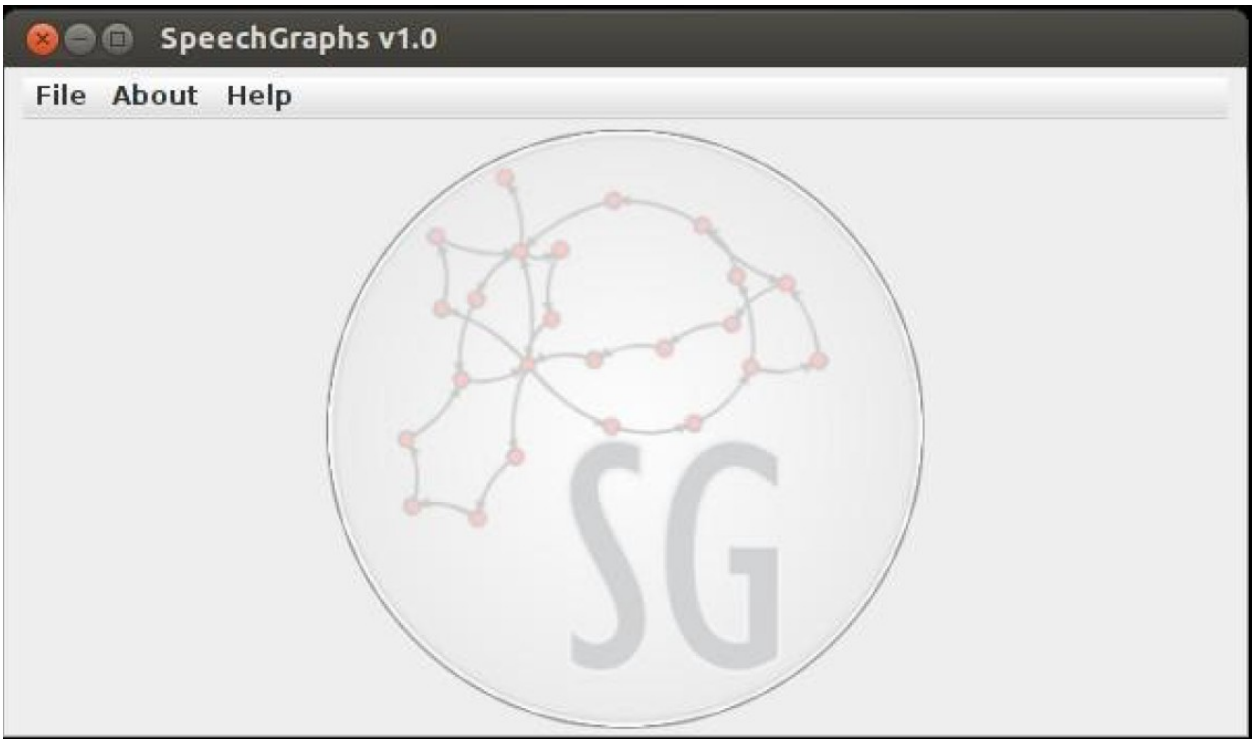
The program represents the loaded text as a graph, i.e. a mathematical representation of a network G = (N, E), with N = {w_1, w_2, ... w_n} that is a set of nodes and E = {w_i, w_j, ... w_n} that is a set of edges. Words are represented as nodes and the temporal links between words are represented as edges. SpeechGraphs generates both a directed graph (digraph) and an undirected graph without parallel or repeated edges, or self-loops (an edge linking a node with itself).

4.1.2 Generating visualization set of data:

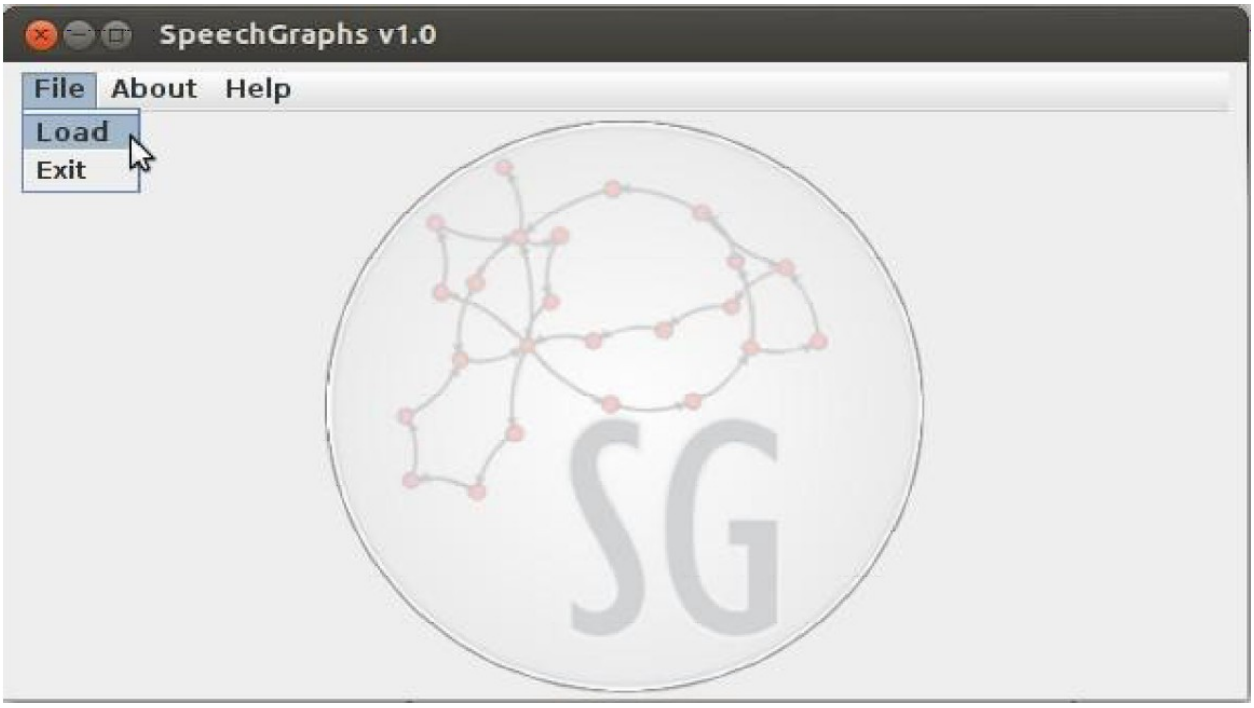
To visualize a set of graph parameters as Graph Plot, Graph Histogram Plot and Degree Distribution Plot for both directed and undirected graph, follow the steps:

1. Open SpeechGraphs, by clicking the jar file on your computer.

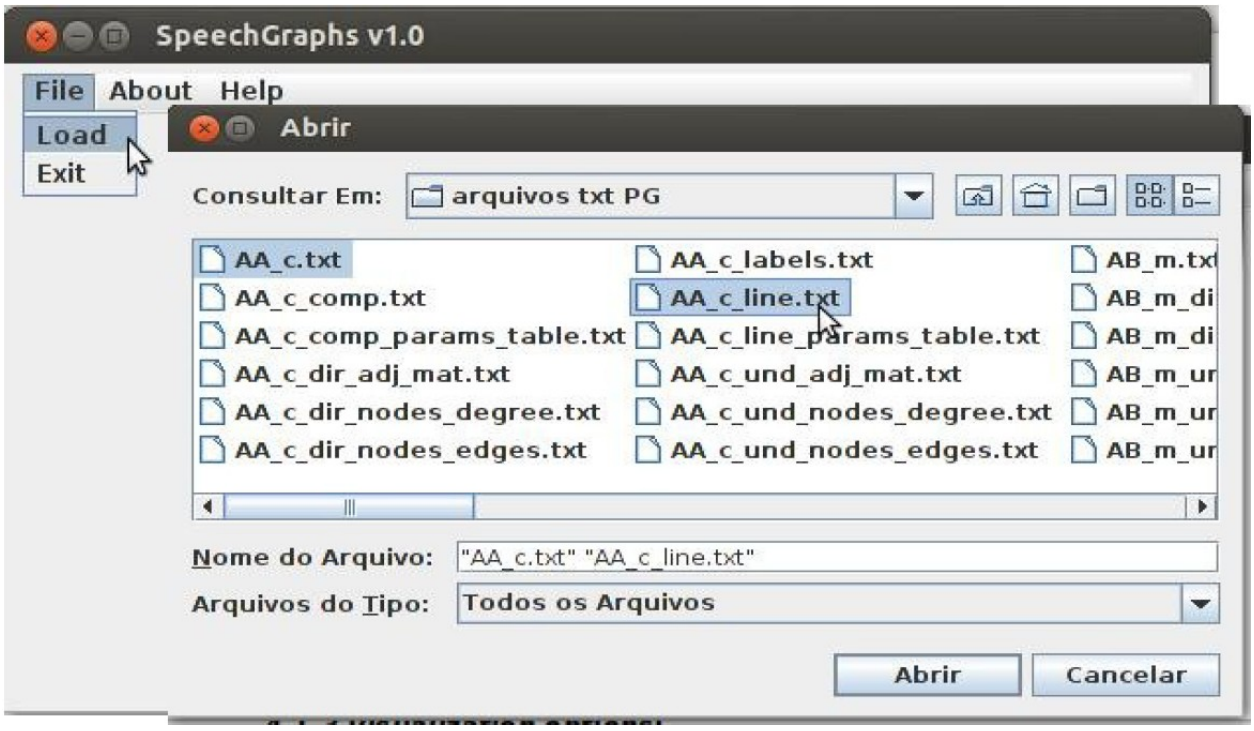
2. The main screens of the application will appear, as shown below:



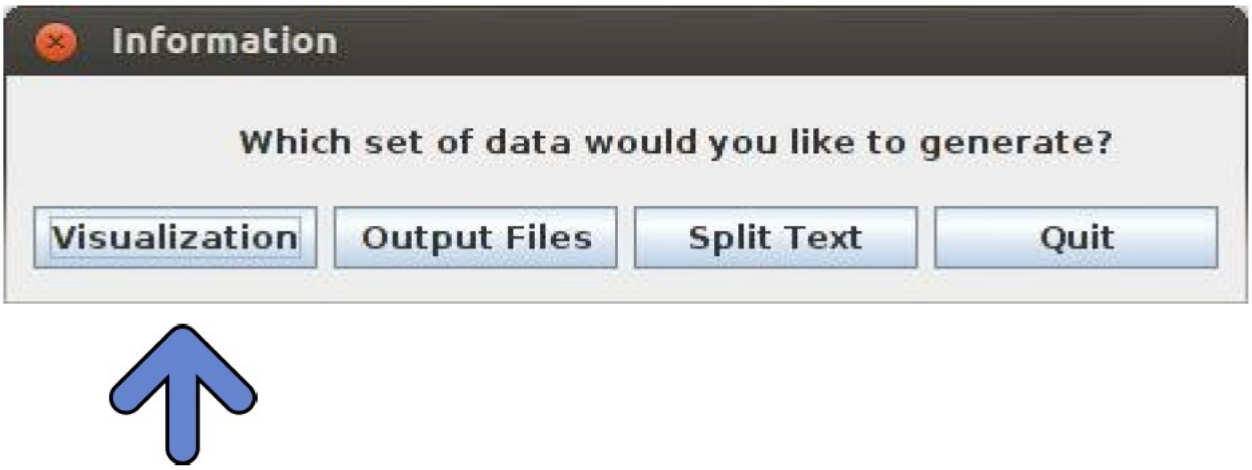
3. To load a file into SpeechGraphs, click on the **File Menu**, Option **Load**, as shown in the picture below:



4. A file chooser window will open. Select the file(s) that you want to analyse:



5. Click on the Visualization button:



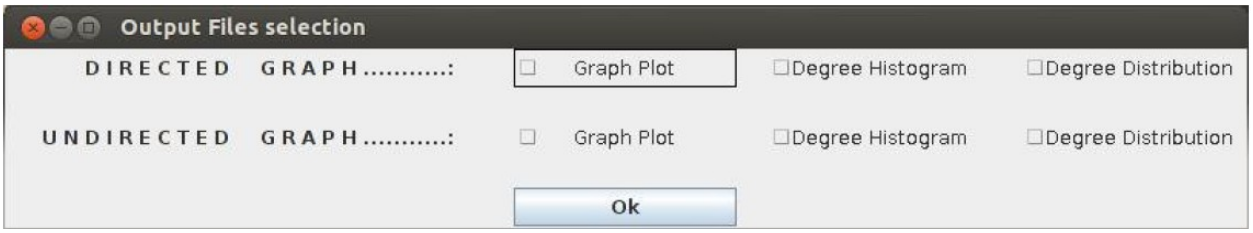
6. After that, you must inform if you want to extract Stop Words, as shown below:

- What are Stop Words?

- Stop Words are any words in the text file that you want to exclude from your analysis.

7. If you chose to extract Stop Words, the screen below will appear. Type the list of words that you want to exclude from your analysis.

7. After that, select the set of features that you want to generate for directed graph, undirected graph or both, as shown in the picture below:



4.1.3 Visualization options:

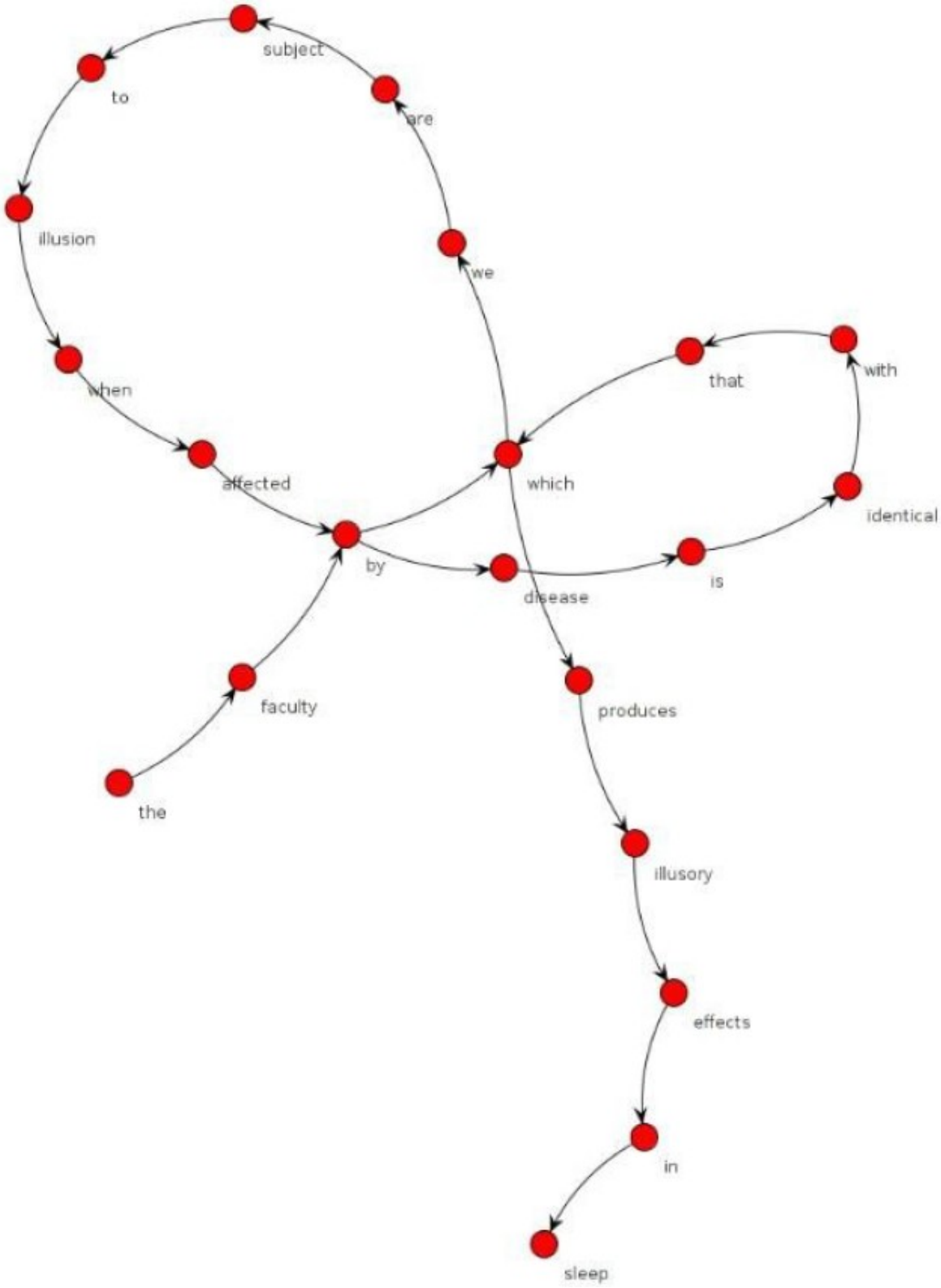
- **Graph Plot** – Shows the plot of the graph(s) selected with its/their parameters. There is an option to export the graph plot to PNG or EPS format (saved on program directory), then click on 'File' (top right corner) and then click 'Export'. There is another option called 'Visualization' that allows the selection of parameters to be visualized on the left window.

- **Node Degree Histogram** – Shows a histogram with the degree of each node in sequence.

- **Degree Distribution** – Shows the number of nodes for each degree value.

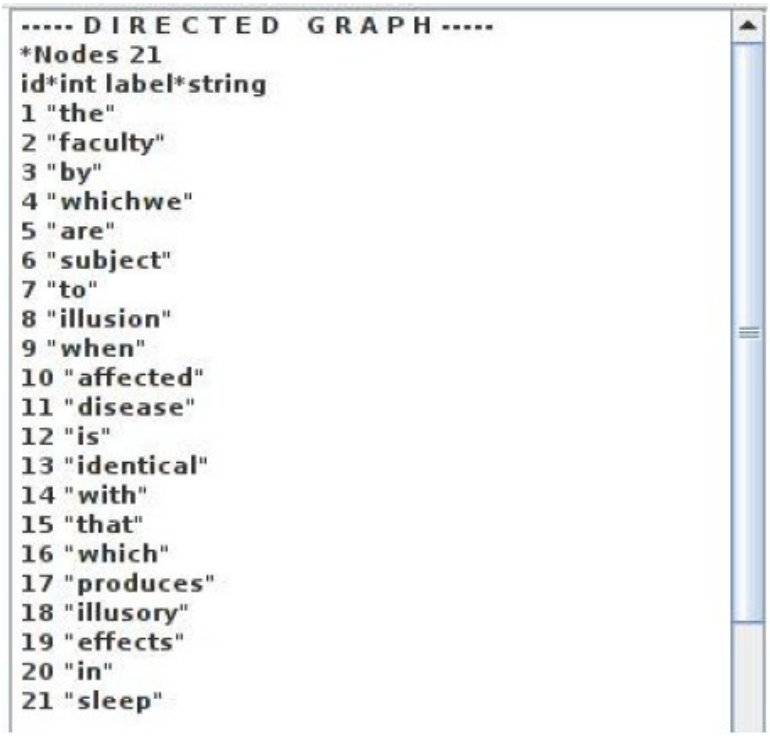
4.1.3.1 Graph Plot window:

If you chose to view the graph plot, you will see the following window:

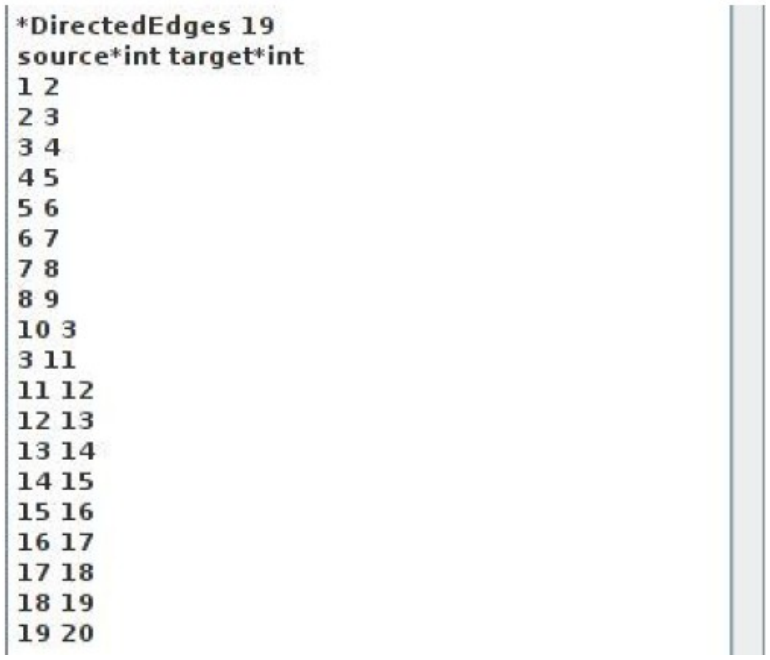

subject
to
are
illusion
we
when
that
with
affected
which
identical
by
disease
is
faculty
produces
the
illusory
effects
in
sleep

On the right side of the window, you will have the following graph parameters:

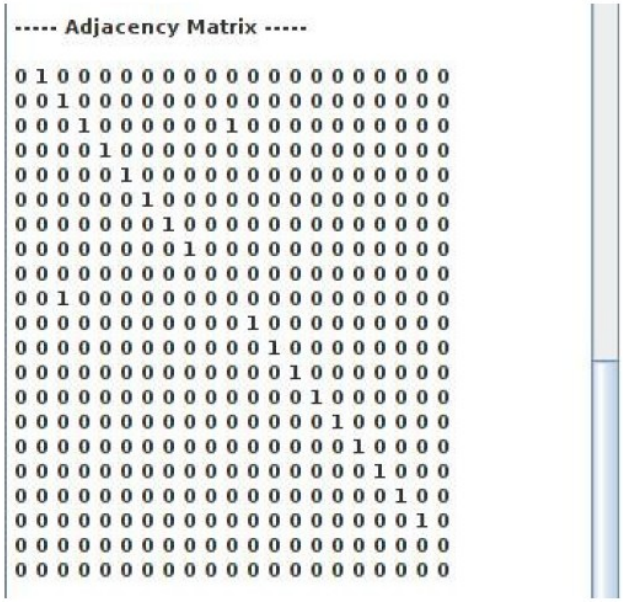
- **Graph nodes list**: describes all the nodes (words) in the text.



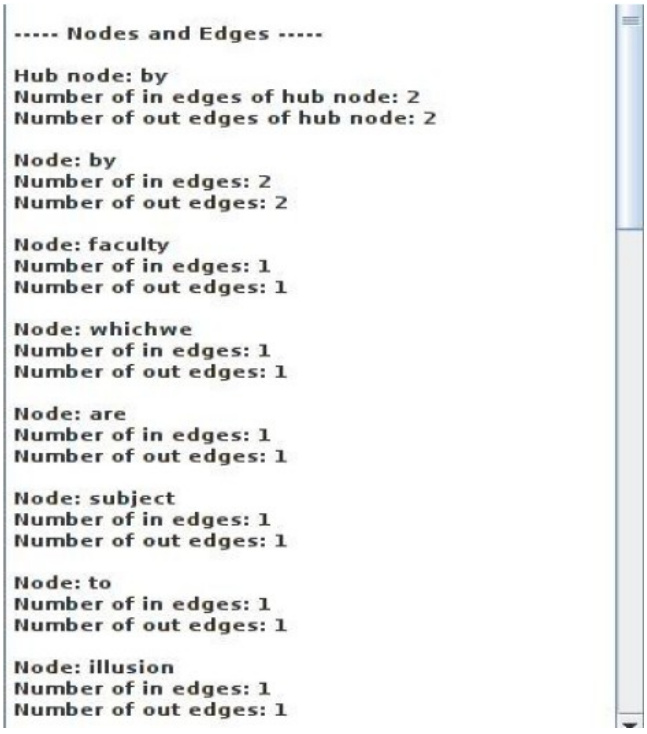
- **Graph edges list**: lists all the edges (links) found between words in the text.



- **Graph adjacency matrix**: for a graph G with n nodes, corresponds to a n x n matrix, considering a_{ij} the number of edges from node i to node j.



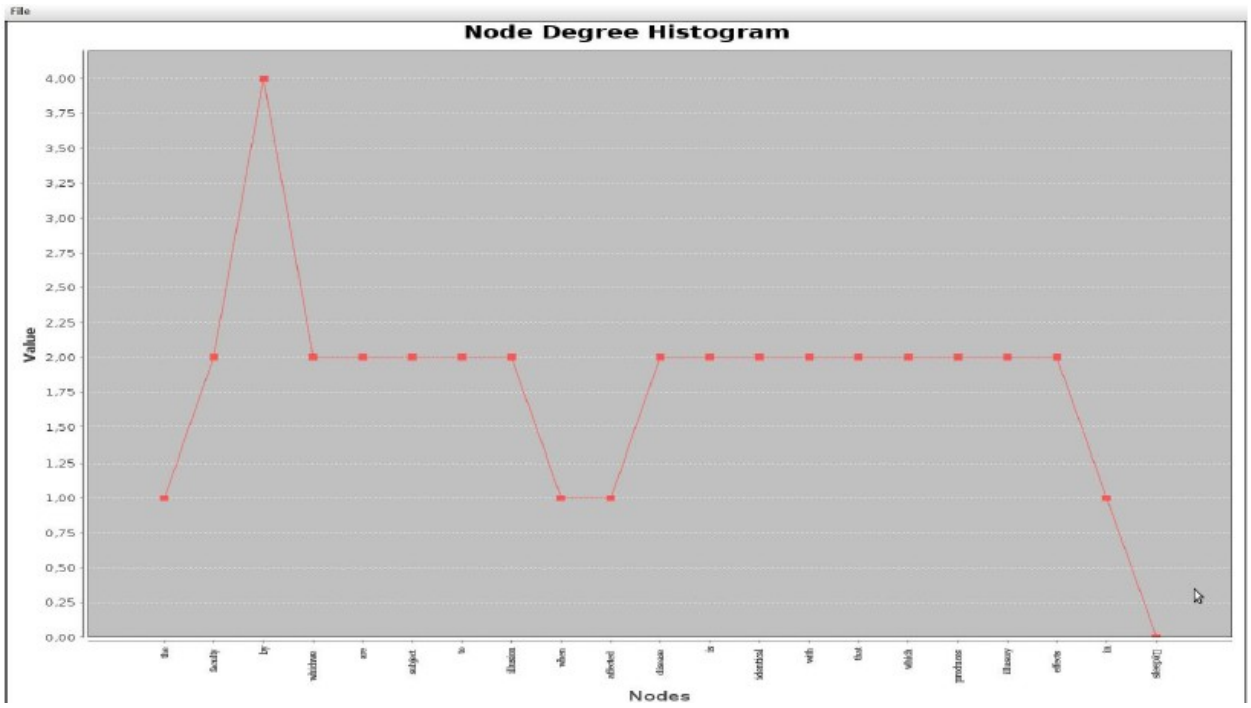
- **Graph hub node**: plots the most connected node (hub); then, lists the nodes according to a decreasing degree sequence; for directed graphs, also plots the number of “in and out” edges of each node.



- **Graph parameters**: description on section: 4.2. Output Files Approach, on Parameters table file.

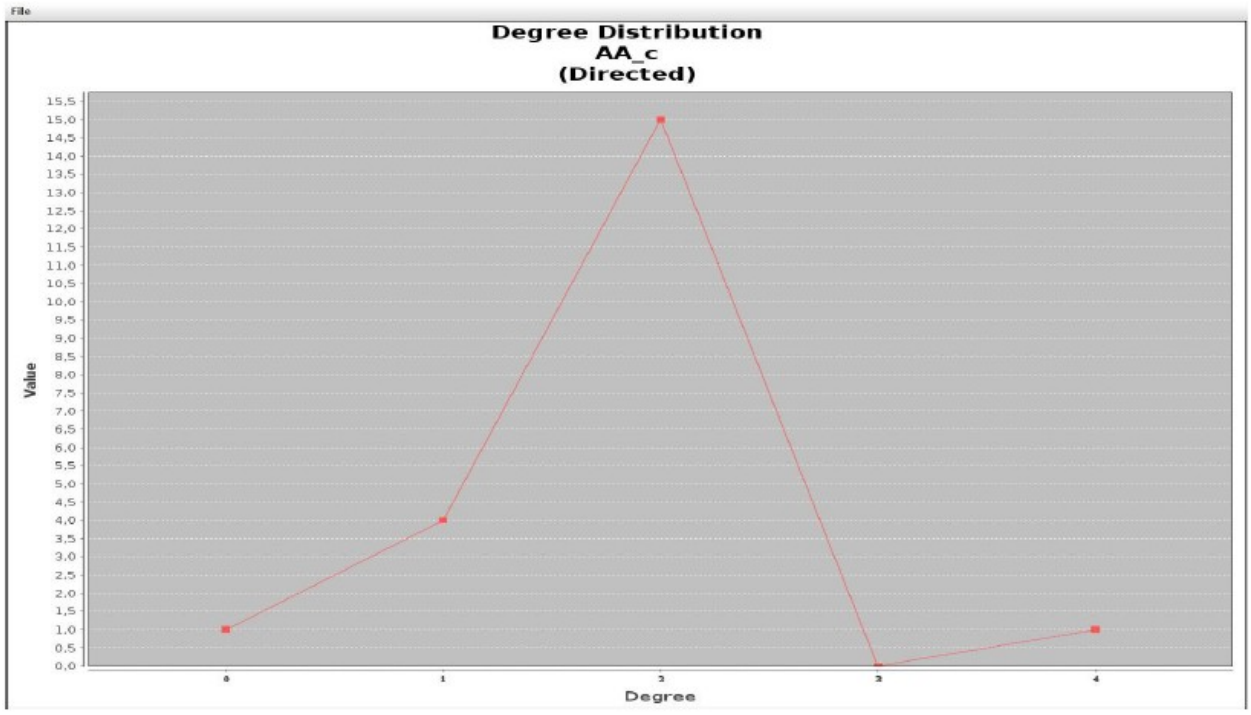
4.1.3.2 Node Degree Histogram:

If you chose to view the Node Degree Histogram, you will see the following window:



4.1.3.4 Degree Distribution:

If you chose to view the Degree Histogram, you will see the following window (X axis you have the degree value and on Y axis you have the number of nodes):



4.2. Output Files Approach

The Output File approach will generate the following files:

- **Nodes and Edges File** – Contains all nodes and all edges of the graph.

Name: <text file name>_<dir/und>_nodes_and_edges.txt

Format:

**Nodes 21*

*id*int label*string*

1 "the"

2 "faculty"

3 "by"

4 "whichwe"

5 "are"

6 "subject"

7 "to"

8 "illusion"

9 "when"

10 "affected"

11 "disease"

...

**DirectedEdges 19*

*source*int target*int*

1 2

2 3

3 4

4 5

5 6

6 7

7 8

...

- **Adjacency Matrix File** – Contains the adjacency matrix of the graph.

Name: <text file name>_<dir/und>_adj_mat.txt

Format:

----- *Adjacency Matrix* -----

0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1

0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0

4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0

- Nodes and Degree File – Contains a list of nodes and their degrees.

Name: <text file name>_<dir/und>_nodes_degree.txt

Format:

Node Degree

7 22

6 16

42 10

24 8

3 6

4 6

25 6

30 6

35 6

...

- Parameters Table File – Contains the table of graph parameters.

Name: params_table.txt

Format:

File, WC, Nodes, Edges, RE, PE, L1, L2, L3, LCC, LSC, ATD, Density, Diameter, ASP, CC

AA_c, 117, 67, 116, 13, 14, 0, 1, 3, 67, 65, 3.4626, 0.0454, 11, 3.9090, 0.07 24

Parameters Description:

File: Name of the text file.

Word Count (WC): Number of words in the text.

N: Number of nodes.

E: Number of edges.

RE (Repeated Edges): sum of all edges linking the same pair of nodes.

PE (Parallel Edges): sum of all parallel edges linking the same pair of nodes given that the source node of an edge could be the target node of the parallel edge."

L1 (Loop of one node): sum of all edges linking a node with itself, calculated as the trace of the adjacency matrix.

L2 (Loop of two nodes): sum of all loops containing two nodes, calculated by the trace of the squared adjacency matrix divided by two.

L3 (Loop of three nodes): sum of all loops containing three nodes (triangles), calculated by the trace of the cubed adjacency matrix divided by three.

LCC (Largest Connected Component): Number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the underlying undirected subgraph.

LSC (Largest Strongly Connected Component): Number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the directed subgraph (node a reaches node b, and b reaches a).

ATD (Average Total Degree): Given a node n, the Total Degree is the sum of “in and out” edges. Average Total Degree is the sum of Total Degree of all nodes divided by the number of nodes.

Density: Number of edges divided by possible edges. (D = 2*E/N*(N-1)), where E is the number of edges and N is the number of nodes. This parameter is based on the undirected graph.

Diameter: Length of the longest shortest path between the node pairs of a network. This parameter is based on the undirected graph. This parameter is based on the undirected graph. If the graph is not connected, this parameter is calculated based on the LCC of the graph.

Average Shortest Path (ASP): Average length of the shortest path between pairs of nodes of a network, calculated by the sum of all shortest path divided by number of paths. If the graph is not connected, this parameter is calculated based on the LCC of the graph.

CC (Average Clustering Coefficient): Given a node n, the Clustering Coefficient Map (CCMap) is the set of fractions of all n neighbours that are also neighbours of each other. Average CC is the sum of the Clustering Coefficients of all nodes in the CCMap divided by number of elements in the CCMap. This parameter is based on the undirected graph.

- LSC File -

Name: <text_file_name>_LSC.txt

Content: Prints the nodes that are part of LSC.

- LCC File -

Name: <text_file_name>_LCC.txt

Content: Prints the nodes that are part of LCC.

- Generate Random Graph: This option allows the user to generate random graphs (same nodes, but randomized set of edges). The field to the right of the option checkbox informs the program how much of such random graphs should be generated.

The generate Random Graph option triggers the output of following files:

- Means parameters File -

Name: means_params_table_random.txt

Format: the same of the params_table.txt

Content: the mean of the params_table.txt file properties, only considering the _random entries.

- Median parameters File -

Name: median_params_table_random.txt

Format: the same of the params_table.txt

Content: the median of the params_table.txt file properties, only considering the _random entries.

- Zscore File -

Name: zscore.txt

Format: the same of the params_table.txt

Content: the zscore calculation based on the graph and its random generated graphs.

- Hub File -

Name: hub_table.txt

Format: File,Hub,Degree,2nd Hub,Degree,Pair,Degree,LSCHub,Degree

4.2.1 Generating Output Files

On the approach menu window, click on the button **Output Files**, as shown below:

The output files will be generated in the same directory of the text file(s).

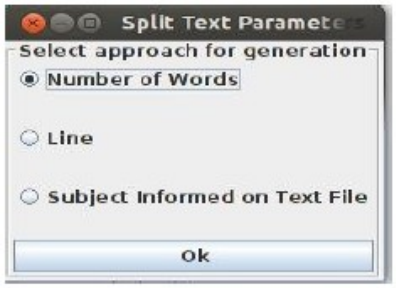
4.3 Split Text Approach

This approach can be used to split the text by a fixed number of words, by text line or by a given subject informed on the text file. As a result, different graphs are created from the same text file.

On the main menu, select the **Split Text** Button:

The Split Text approach is divided into three methods:

1. **Number of Words**: This option allows for a moving window analysis. You must inform the bin size (number of words per graph) and the step size (numbers of words to be skipped to create the next graph), as shown in the picture below:


Split Text Parameter
Select approach for generation
Number of Words
Line
Subject Informed on Text File
Ok


Input
Inform the Bin Size (number of words)
OK
Cancel


Input
Inform the Step
OK
Cancel

How bin size and step work?

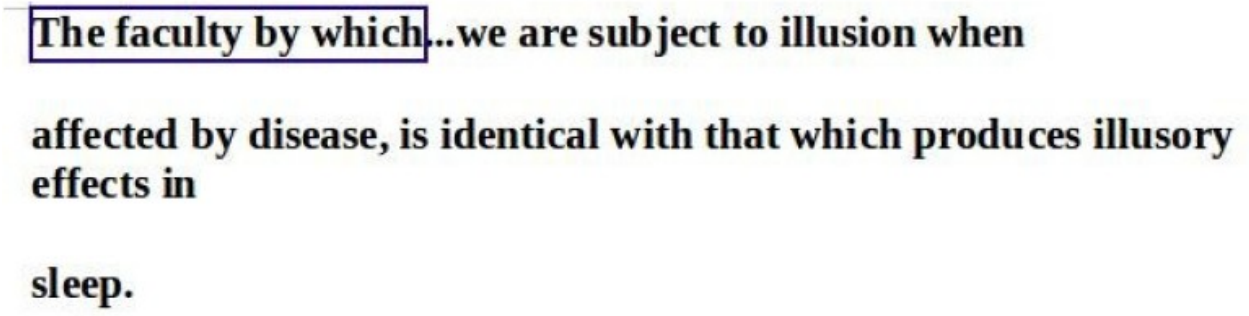
In this approach we will have a sliding window through the whole text. The window will forward depending on the bin size and step size. To illustrate, Let's take the transcription example below, with a bin size of 4 (four) words and step of 3 (three) words.

The faculty by which...we are subject to illusion when

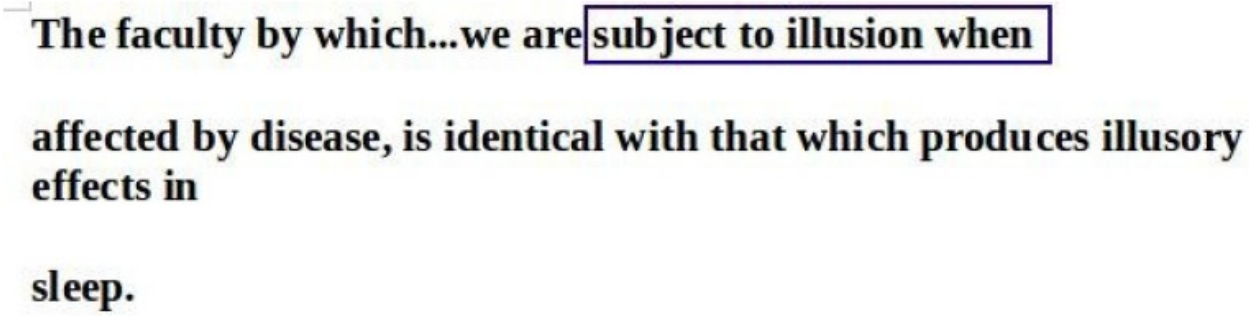
affected by disease, is identical with that which produces illusory effects in

sleep.

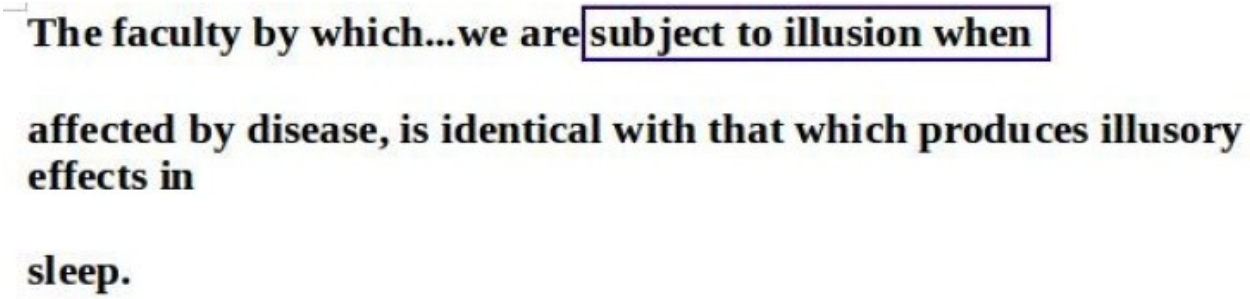
1st Iteraction



2nd Iteraction



3rd Iteraction



4th Iteraction

The faculty by which...we are subject to illusion when

affected by disease, is identical with that which produces illusory effects in

sleep.

5th Iteraction

The faculty by which...we are subject to illusion when

affected by disease, is identical with that which produces illusory effects in

sleep.

6th Iteraction

The faculty by which...we are subject to illusion when

affected by disease, is identical with that which produces illusory effects in

sleep.

7th Iteraction

The faculty by which...we are subject to illusion when

affected by disease, is identical with that which produces illusory effects in

sleep.

8th Iteraction

The faculty by which...we are subject to illusion when affected by disease, is identical with that which produces illusory effects in sleep.

The Split Text approach returns two files:

• Parameters table: contains the parameters for each graph generated by a text file.

• Mean Parameters table: contains the mean parameters of all graphs generated by a text file.

2. **Line**: For this option, one graph is generated for each text line. The parameters table contains a line of graph parameters for each line of the text file(s).

3. **Subject Informed on Text File**: For this option, the text file(s) must have the subjects delimited by the tags #{.... #}, as shown below:

#{Dream

A dream is a short-lasting psychosis, and a psychosis is a long-lasting

dream.

#}

With this option you will also have the opportunity to analyse a given subject graph as if all its separate components were together in a single text line, as shown below:

The parameters table contains a line of parameters for each subject graph of the text file(s).

5. Considerations about SpeechGraphs

a) Every word is transformed to lower case in the end of the processing phase, which means that “Dream” and “dream” are considered the same word in the graph.

b) The stop words are removed before the transformation to lower case. Because of that, if you type, for instance, the upper case version of the word “a” (“A”) in the stop words dialog box, only the upper case version will be removed “A”. If you also want to remove the lower case version of the word “a” (“a”), you need to add that in the dialog box. So, if you want to remove all occurrences of the word “a” in both lower case and upper case, you have to type “A a” into the dialog box.

c) The following characters are removed in the processing phase: “,”, “.”, “:”, “?”, “!”, “...”, “-”, “;”, “(“, “)”, “>”, “<”, “_”, “=”, “+”, “*”, “$”, “%”, “@”, “&”, “[“, and “]”. If there are, for instance, the word “Dream” and the word “Dream?”, they are both considered the same word, because the special character “?” is removed in the processing phase.

d) It is expected that the input text is written obeying the same spacing rules we use when writing a text. For instance, if you have the sentence: “This is a dream, but I’m not a dreamer.”, you should make sure that the spacing between words and special characters are exactly as shown in the sentence, and not something such as: “dream ,” or “I’ m”, so that the text isn’t improperly processed.

6. Contact

Site: Brain Institute UFRN (http://www.neuro.ufrn.br).

Email: speechgraphs@neuro.ufrn.br.