Hacettepe University
Department of Computer Engineering
BBM 204
Programming Assignment 4

May 13, 2013

Subject: Minimum Spanning Trees & Shortest Path in Graphs
Submission date: 13th May 2013
Deadline: 27th May 2013 - 12:59pm
Programming Languge: Java
Advisors: Dr. Burcu Can, Dr. Erkut Erdem

1 Introduction

Graphs are widely used in Computer Science for a wide range of subjects. These subjects may include Natural Language Processing, Image Processing, Pattern Recognition, Speech Recognition, Bioinformatics etc.

Graphs provide a natural way of constructing relationships between objects; thereby it is a natural way of learning by using this relationship between objects. These objects may be genes in a dna, phonemes in a speech signal, pixels in an image, words in a text (see Figure 1), and so on. Depending on the type of the problem, everything can be modelled by using the graph theory. There are many operations defined on graphs. You will practice Shortest Path and Minimum Spanning Trees in this assignment:

Shortest path between two vertices in a graph is the shortest way of reaching from one vertex to another. Shortest path is mainly used for finding a relationship between two vertices. Vertices which bear similar items are located close to each other in graphs. Therefore, shortest paths provide a way of measuring how similar two items are.

A Minimum Spanning Tree is a subgraph of a given graph which has the minimum sum of the weights on the edges in the spanning tree. Minimum Spanning Trees have a wide range of application fields. For example, they are

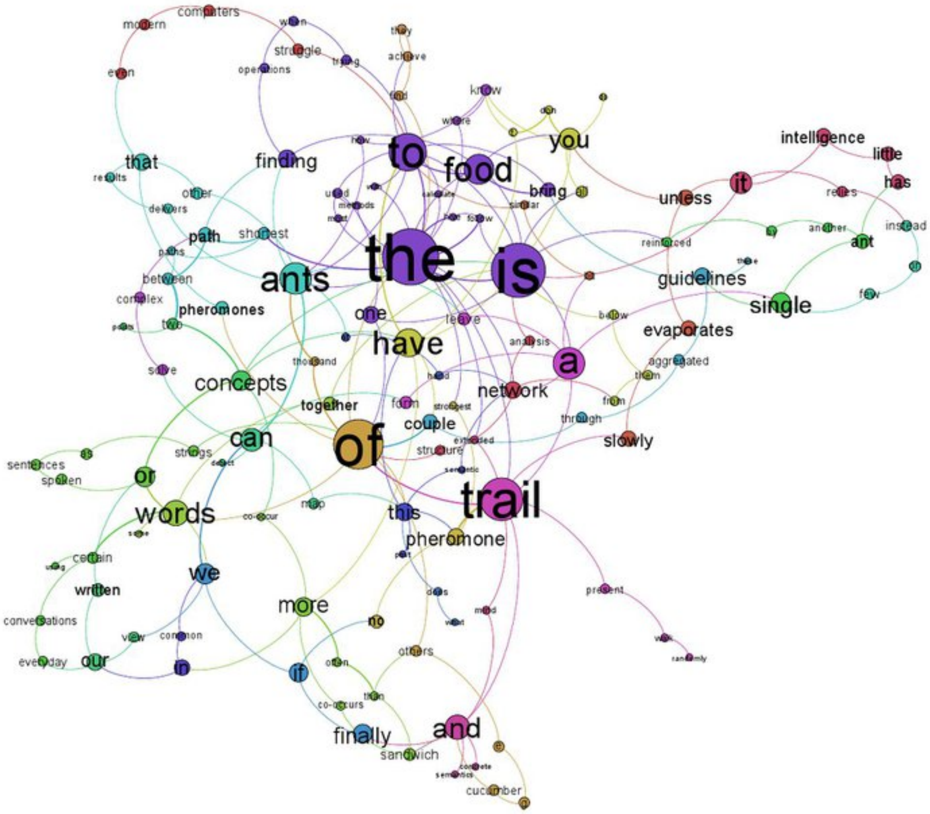


Figure 1: A word graph [3]

widely used in the design of the computer networks or telecommunication networks, handwriting recognition, image segmentation, clustering (see Figure 2), and so on.

2 Definition of the Problem

In this programming assignment, you will practice on one application field of graphs. The field that you will apply graphs is Natural Language Processing. You are expected to measure semantic similarity between words by using the shortest path. Subsequently, you will find word clusters that bear semantically similar words in each of them.

Semantic similarity is a concept that measure how similar two words/documents/terms/concepts/senses are in meaning. For example, words 'blue', 'red', 'yellow' are semantically similar, whereas 'book' is not

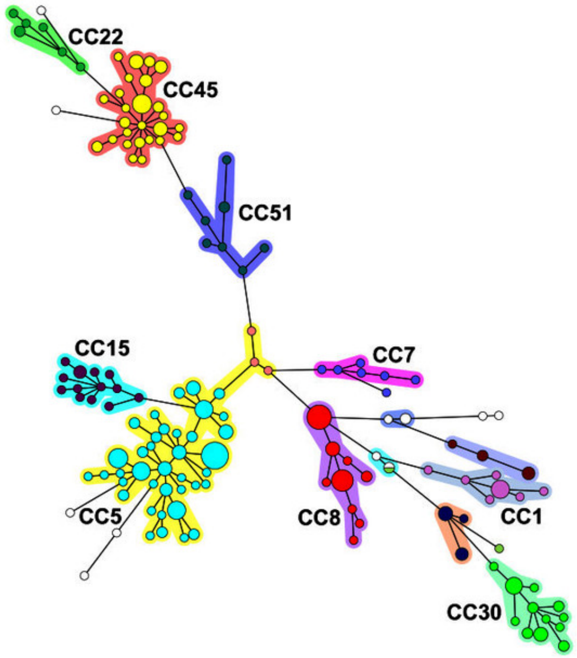


Figure 2: An example of minimum spanning tree clustering [4]

semantically similar to these words. However, book is semantically similar to 'notebook'. In Figure 3, a graph which is constructed by the Flickr tags is given (see Figure 4 for visual images for the given tags). You can see that semantically similar words are located close to each other on the graph, whereas semantically different words are quite distant from each other.

Another point that you need to observe on the graph is that, vertices in different colours refer to different clusters. Moreover, these clusters bear semantically similar words.

In this assignment, you will use a dictionary to build an undirected dictionary graph. Each word in the dictionary is a vertex of the graph. There is an edge from u to v if v appears in the definition or vice versa. For example, if *animal* appears in the definition of *cat*, then an edge will be created between the words *animal* and *cat*. Your initial graph will be an unweighted graph, where each edge has the same weight (i.e. weight=1 for each edge).

The dictionary may consist of various forms of words. These forms include plural forms, tenses etc. For example, if the word *animal* exists in the dictionary, the plural form *animals* may also exist in the dictionary. For

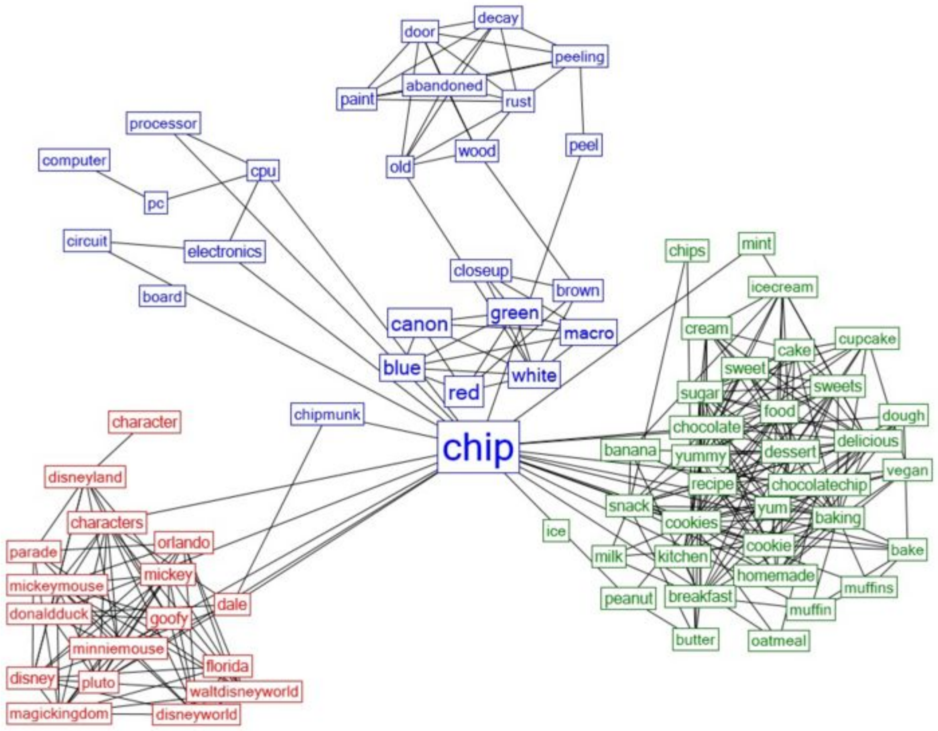


Figure 3: Disambiguation of Flickr Tags [1]

example, if the word *take* exists in the dictionary, various forms of the same word such as *takes*, *taken*, *took*, *taking* may also exist in the dictionary. Please keep in mind that all different forms of the words must belong to different vertices on the graph.

Words must be handled by ignoring the case. For example, the words *animal* and *Animal* will be considered as the same word.

Once the initial graph is constructed as described, you will use your graph for practicing two operations: a) measuring semantic similarity, b) finding clusters.

2.1 Measuring the Semantic Similarity Between Words

You will use your graph to measure semantic similarity between words. There are several ways to measure semantic similarity between words on a graph. However, you will use the edge-based measure [2]:

$$sim(w_1, w_2) = \frac{1}{len(w_1, w_2)} \tag{1}$$

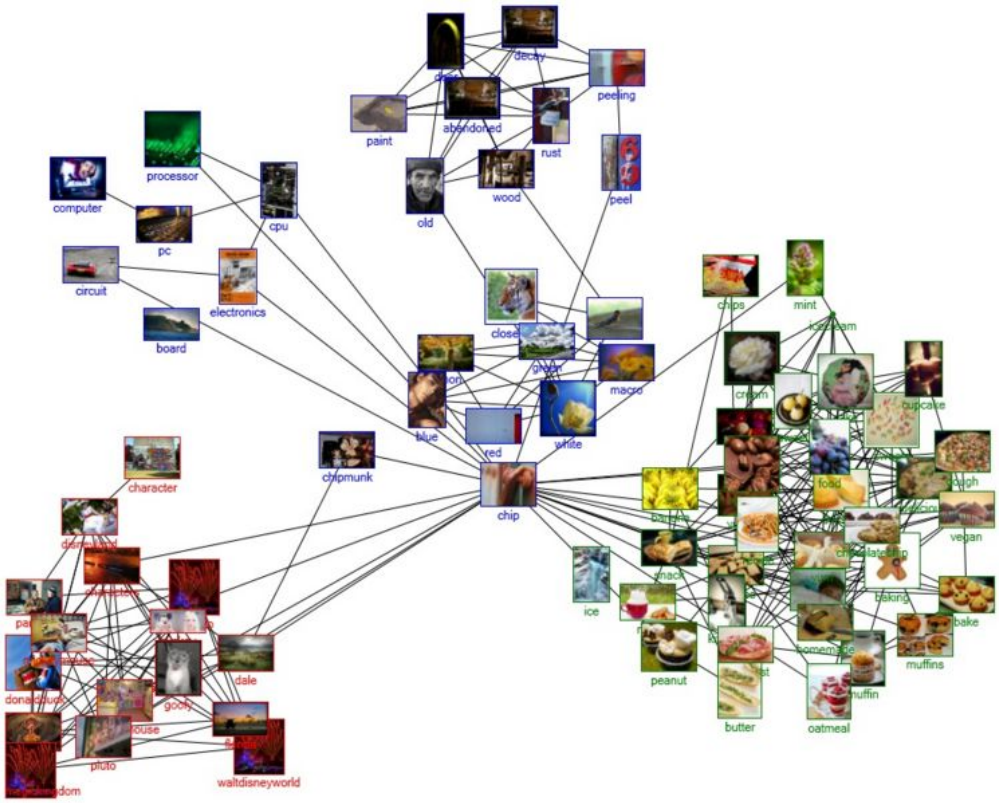


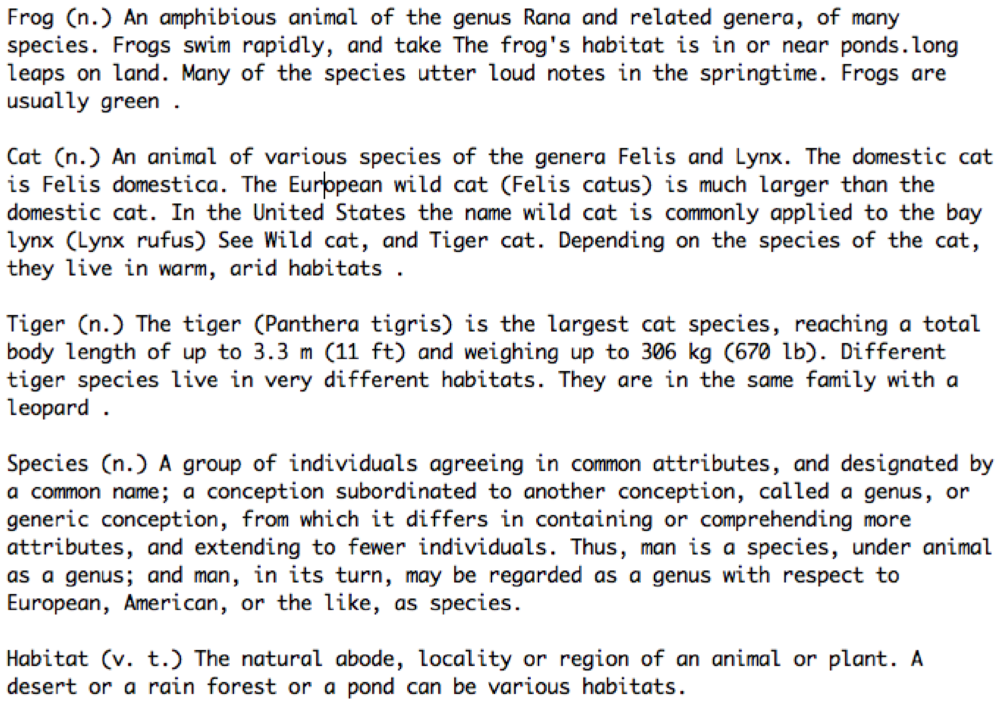
Figure 4: Disambiguation of Flickr Tags [1]

where *len* function is the simple calculation of the shortest path length (i.e. weight=1 for each edge).

2.2 Clustering

In this part of the assignment, you will use minimum spanning trees on your graph to find word clusters (see Figure 2). As indicated before, minimum spanning trees can be used for clustering data. Once the minimum spanning tree on a given graph is found, k clusters can be obtained by removing $k-1$ edges that have the minimum weights on the minimum spanning tree. Therefore, k clusters are obtained. Each of these clusters are expected to have semantically similar words.

Since your initial graph is not weighted, you will construct a weighted graph before finding the minimum spanning tree. In order to create a weighted graph, you will remove all the edges from your initial graph, and you will add only the edges between the word pairs that you measured the semantic



```
Frog (n.) An amphibious animal of the genus Rana and related genera, of many
species. Frogs swim rapidly, and take The frog's habitat is in or near ponds.long
leaps on land. Many of the species utter loud notes in the springtime. Frogs are
usually green .

Cat (n.) An animal of various species of the genera Felis and Lynx. The domestic cat
is Felis domestica. The European wild cat (Felis catus) is much larger than the
domestic cat. In the United States the name wild cat is commonly applied to the bay
lynx (Lynx rufus) See Wild cat, and Tiger cat. Depending on the species of the cat,
they live in warm, arid habitats .

Tiger (n.) The tiger (Panthera tigris) is the largest cat species, reaching a total
body length of up to 3.3 m (11 ft) and weighing up to 306 kg (670 lb). Different
tiger species live in very different habitats. They are in the same family with a
leopard .

Species (n.) A group of individuals agreeing in common attributes, and designated by
a common name; a conception subordinated to another conception, called a genus, or
generic conception, from which it differs in containing or comprehending more
attributes, and extending to fewer individuals. Thus, man is a species, under animal
as a genus; and man, in its turn, may be regarded as a genus with respect to
European, American, or the like, as species.

Habitat (v. t.) The natural abode, locality or region of an animal or plant. A
desert or a rain forest or a pond can be various habitats.
```

Figure 5: A sample dictionary

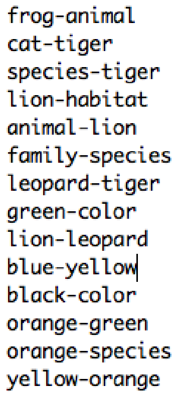
similarities in the first step. Once you have a weighted graph, you can find the minimum spanning tree and cut $k-1$ edges that have the minimum weights on the graph to obtain k clusters.

The number of clusters will be provided as a parameter. Therefore, the number of clusters to be obtained will be fixed initially.

3 Input-Output & Testing

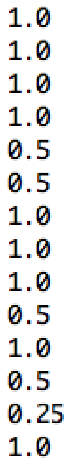
You will have two input files and two output files in this assignment. All of the file names will be provided to the program as command line arguments.

- **dictionary file:** A dictionary is given in this file. Every word and its definition are given in different lines. A sample dictionary is given in Figure 5.
- **word pairs file:** A word pairs list is given in this file. Your program is expected to measure the semantic similarities between the word pairs. A sample word-pairs file is given in Figure 6.



```
frog-animal
cat-tiger
species-tiger
lion-habitat
animal-lion
family-species
leopard-tiger
green-color
lion-leopard
blue-yellow
black-color
orange-green
orange-species
yellow-orange
```

Figure 6: A sample word-pairs file

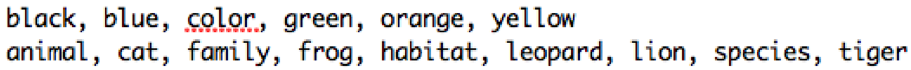


```
1.0
1.0
1.0
1.0
0.5
0.5
1.0
1.0
1.0
0.5
1.0
0.5
0.25
1.0
```

Figure 7: A sample semantic similarities file

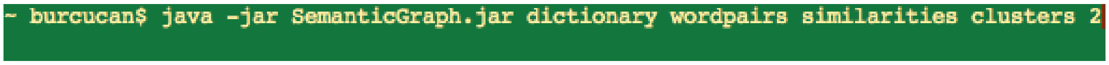
- **semantic similarities file:** The results of the semantic similarities between word pairs will be written to this file. A sample semantic similarities file is given in Figure 7.

- **clusters file:** Finally, your program will produce the minimum spanning tree and produce the clusters out of this spanning tree. Contents of each cluster will be written to this file. Cluster members will be written in alphabetical order (in increasing order) and clusters will be written according to the number of members in each cluster in increasing order (i.e. cluster that has the minimum number of members will be written first). Cluster members must be delimited by commas. A sample clusters file is given in Figure 8.



```
black, blue, color, green, orange, yellow
animal, cat, family, frog, habitat, leopard, lion, species, tiger
```

Figure 8: A sample clusters file



```
~ burcucan$ java -jar SemanticGraph.jar dictionary wordpairs similarities clusters 2
```

Figure 9: A terminal screenshot that shows how your program will be run from the command line.

4 Execution of the Program

Your program will be run from the command line as follows:

java -jar SemanticGraph.jar dictionary wordpairs similarities clusters numberofclusters
Here:

- *dictionary* is the name of the file that has the dictionary,
- *wordpairs* is the name of the file that has word pairs that your program will measure how similar they are,
- *similarities* is the name of the file that your program will write the semantic similarities of word pairs in the *word pairs* file,
- *clusters* is the name of the file that the contents of each cluster will be written to,
- *numberofclusters* is the number of clusters that your program has to find by using the minimum spanning tree.

Please keep in mind that all of the names of the files will be taken as command line arguments and will not be fixed names, otherwise your program will not be able to run with other file names.

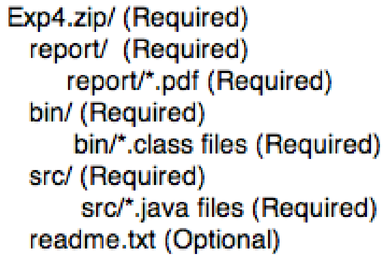
A screenshot of the terminal that shows how to run your program is given in Figure 9.

You can test your program by creating your own dictionary. You can use the plain text English dictionary given in:
`http://www.mso.anu.edu.au/~ralph/OPTED/` to test your program.

You can also test your program for other languages (i.e. Turkish, Spanish etc.). In addition, you can also test your program by related Flickr tags. In that case, each line will consist of a tag and related tags. You do not need to change anything in your implementation to test your program with Flickr tags.

5 Submission

Your submission format must be as given in Figure 10.



```
Exp4.zip/ (Required)
  report/  (Required)
     report/*.pdf (Required)
  bin/ (Required)
      bin/*.class files (Required)
  src/ (Required)
       src/*.java files (Required)
  readme.txt (Optional)
```

Figure 10: Submission format

6 Notes and Restrictions

- The very best advice I can give you is: do not wait until the last minute!
- Your experiment should be submitted before the due date. Late submissions will be penalised.
- Keep in mind that design and programming are individual creative processes!
- You must specifically describe in your readme.txt file, whatever help (if any) that you received from others and tell us the names of any individuals with whom you collaborated. This includes help from friends, classmates, lab teaching assistants, course staff members and the web.
- Do not, under any circumstances, copy another person's code. Incorporating someone else's code into your program in any form is a violation of academic regulations. This includes adapting solutions or partial solutions to assignments from any offering of this course or any other course.
- Explain the data structures and the algorithms that you used in your implementation clearly in your report.

References

[1] Hyunjong Cho and Viet-An Nguyen. Flickr tags disambiguation, February 2011.

[2] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *in Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997.

[3] Olli Parviainen. Shortest path between two concepts is the most used one, February 2011.

[4] Hoang Vu-Thien, Katia Hormigos, Gaelle Corbineau, Brigitte Fauroux, Harriet Corvol, Didier Moissenet, Gilles Vergnaud, and Christine Pourcel. Longitudinal survey of staphylococcus aureus in cystic fibrosis patients using a multiple-locus variable-number of tandem-repeats analysis method. *BMC Microbiology*, 10(1), 2010.