Assignment 4 – CIR(Viz)

Submission Date: October 23, Monday *Report(by 10pm)*
Demo(During Tutorial hrs in the week of Oct 23. Schedule to be posted near to the time)

Type: Pair work **Total marks of the assignment: 20** **Weightage: 5%**

I. Introduction

We understand and retain information better when it is visually presented. With our decreasing attention span (~8 minutes), and a constant exposure to information, it is crucial that we communicate information in a quick and visual way. **Patterns** or **insights** which often go unnoticed in a dataset, become obvious if we put the same information on say a graphical chart. One of the key benefits of data visualization is how it enables users to more effectively see connections between diverse data points. In today's highly complex systems, finding these correlations among the data has never been more important. Data visualization also allows us to quickly interpret the data and adjust different variables to see their effect. Visualization Tools e.g. Tableau, D3, are increasingly making it easier for us to do so.

The purpose of the assignment is to:

a) Build on Assignment 3. Extract first 200,000 lines from the (full) dataset at http://labs.semanticscholar.org/corpus/ and use them for the visualization tasks listed in section III. Note that the dataset is ~10GB in size. The download shall take some time. Also the 200,000 lines will require a few GB of RAM for processing. Please contact us at cs3219.cir@gmail.com in case you face any problem with download or extraction of data.

b) Learn a new skill of visualization and a tool e.g. d3.js. (Tutorial 6 has information on d3. See IVLE Files- Tutorial folder). However, you could choose any visualization tool or framework to do this assignment.

II. Important Information

1. Read this document carefully.

2. Use any tool or language to address the requirements. You can use any resources e.g. online tutorials/ textbooks.

3. If you have any query about this Assignment, send mail to **cs3219.cir@gmail.com**

4. Use first 200,000 lines of (full) dataset http://labs.semanticscholar.org/corpus/ for this assignment. Note that the dataset file is about 7GB in size. The extraction of data may take several mins. Also the 200,000 lines will require about a few GB of RAM for processing. Please

contact us at cs3219.cir@gmail.com or drop a note to lecturer in case you have any issues with extraction time or RAM size of your machine.
A note about dataset:

The dataset at http://labs.semanticscholar.org/corpus/ provides data about over 7 million published research papers in Computer Science and Neuroscience.

You will find two links – Full and Sample.

For the assignment, extract first 200,000 lines of the FULL dataset.

The link also gives short description of data attributes and an example.

5. Demonstrate visualizations, corresponding to the tasks set in Section III, **to your tutor on Monday, October 23, during tutorial hrs** (a schedule will be posted near to the time). *Note there will not be any regular tutorial on Monday 23 October or on Wednesday 25 October.*

6. Submit a report (1-3 pages, single pdf), **as per Report template given at the end of this document, by Monday, October 23, 10PM** in IVLE folder A4-CIRViz-Report in IVLE Files(workbin). Exceeding the page guideline of 1-3 pages does not invite any penalty.
Label the report document: A4_<Matric-number-1>_<Matric-number-2>
e.g. A4_A0045396X_A0046342Y.pdf

III. Task

Use **2 or more different types of visualizations to achieve the tasks given below**. **Each task should be covered by at least 1 visualization**.

There are in total **5 tasks:**

1. Visualize **the top 10 authors for venue arXiv** based on **the number of publications he/she has made across all available years for arXiv.**

2. Visualize the **top 5 papers for venue arXiv** based on **the number of citations across all available years for arXiv.** (how many times this paper has been cited, so consider those with the largest inCitations from **arXiv)**

3. Visualize the trend of the amount of publications across all available years for **venue ICSE**.

4. Citation makes up a major proportion of information researchers will use while analyzing the scientific publication dataset. Construct a **citation web for the base paper with title " Low-density parity check codes over GF(q)"** . You could create a

visualization to illustrate up to 2 levels of base paper citation i.e. if the base paper is A, capture up to C: A is cited by B, and B is cited by C, so A <- B <- C.

The base paper and citation paper should be displayed in different colors for distinguishing purpose; a line should be explicitly constructed linking the base paper and the citation paper; for each paper, display its available relevant information e.g. its title and authors.

5. Create a Visualization of your own choice, based on any other relevant query of your choice (e.g. you could take a query from Assignment 3)

Note:

a) you can use any type(s) of visualization as long as you can achieve the above tasks. We value creativity.

b) You may need to do extra processing on data before visualization. You don't have to report the extra processing script.

c) Before you start, take a look at the sample dataset provided in the link and get a sense of how the actual data looks like. Basically the two have the same format; the actual data is only bigger in size.

d) When constructing the citation web, you'll need to find those papers which cites the base paper; however, as you'll only extract the first 200,000 lines of the raw dataset, some information will be missing: the citation is established using ID of the papers, so let's say there are 300 papers cite base paper A, it's possible that you can only find titles of 15 of them. In that case, include only those 15 papers. We need to see the titles and authors in the graph, so if the dataset doesn't contain it, do not include it.

e) To help you get a better understanding on what you need to do, we provide a few visualizations, in Appendix I, based on the dataset used in A3 (the XML dump).

d) Here are some links for different types of visualizations.

A. Pie chart

Sample visualization: https://bl.ocks.org/mbostock/3887235

B. Heat map

Sample visualization: http://bl.ocks.org/tjdecke/5558084

C. Stacked bar chart

Sample visualization: https://bl.ocks.org/mbostock/1134768

D. Google-calendar-like visualization

Sample visualization: http://bl.ocks.org/chaitanyagurrapu/6007521

E. Waterfall Chart

Sample visualization: http://bl.ocks.org/chucklam/f3c7b3e3709a0afd5d57

F. Bubble Chart

Sample visualization : https://bl.ocks.org/mbostock/4063269

Report template on next page.

--- report template ---

Assignment 4: CIR (VIz)

Student Name		
Matriculation Number		

1. Introduction

(up to 1 paragraph including objective of assignment in your own words; individual contribution of each member in doing this assignment)

2. Visualizations - Purpose & Method

(i) State which visualization(s) did you select for each of the objectives given in Section III. In order to facilitate grading, you can use a table showing which objectives are covered by which visualization. *An example is given below:*

Objective	*Visualization*
1	*Heatmap*
2	*PieChart,* *Heatmap*
3	*Waterfall*

(ii) Provide an image of each of the visualizations you created

(iii) For any one of the visualizations:

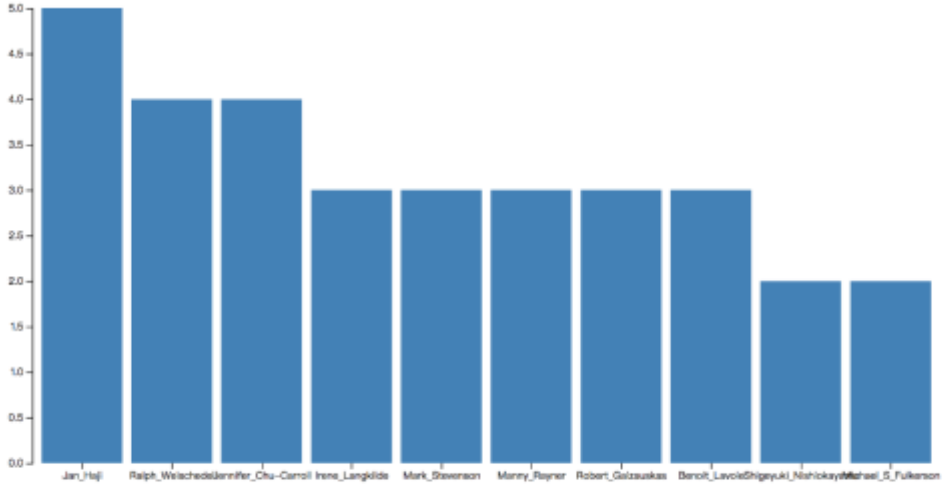
List Step wise method you followed in creating the visualization. Be precise and succinct . Include CIR APIs from your A3 submission in case you used any. Write with a perspective such that your peers could easily use your method to create a similar visualization.

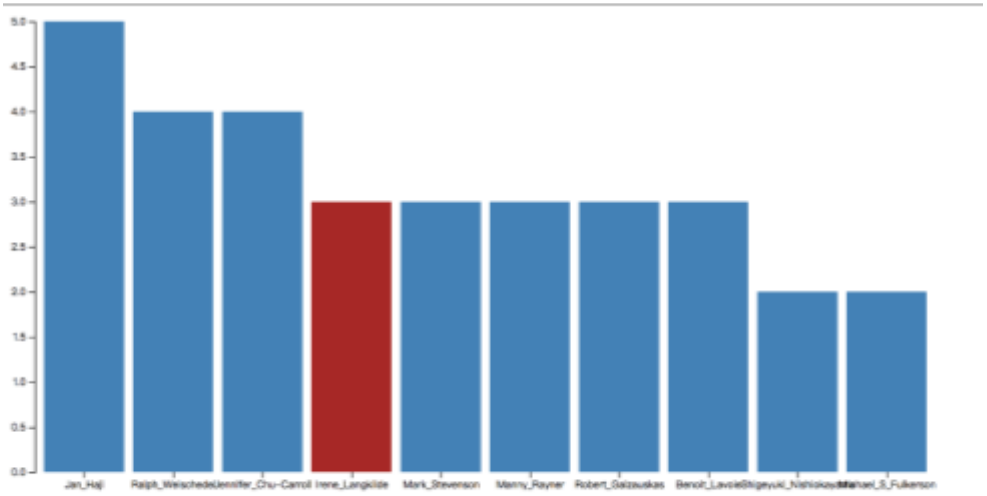
3. (optional) Any other comments or information you may have

Appendix I

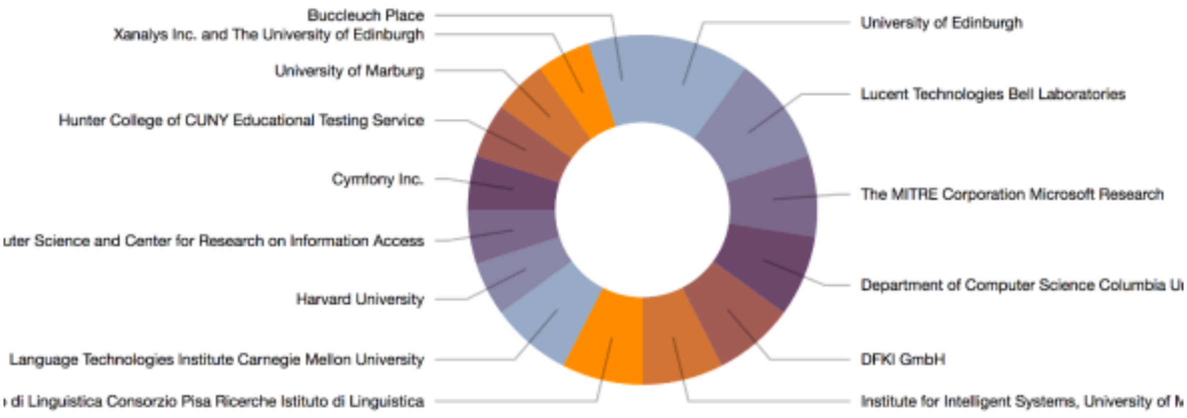
The graph/plots given below are based on the data from folder A00.

1. The top 10 authors:

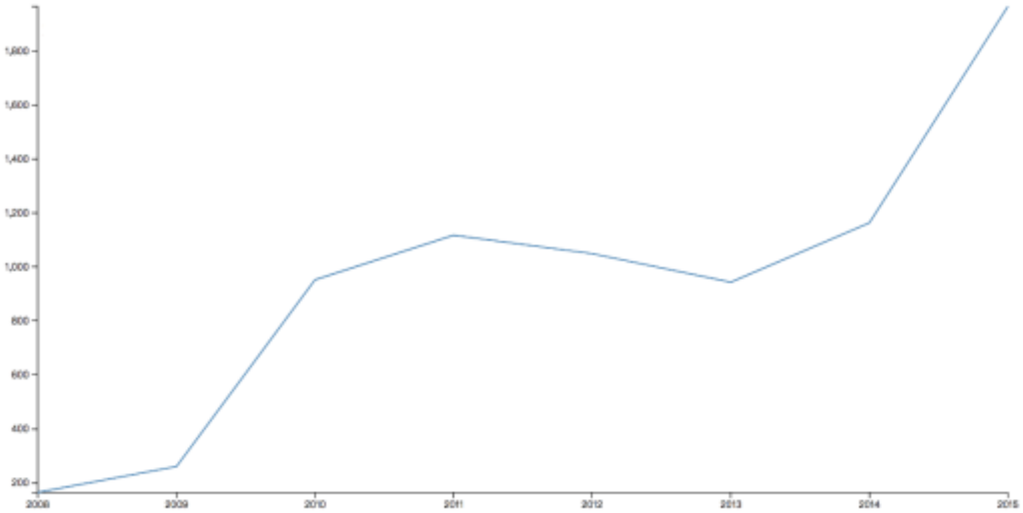




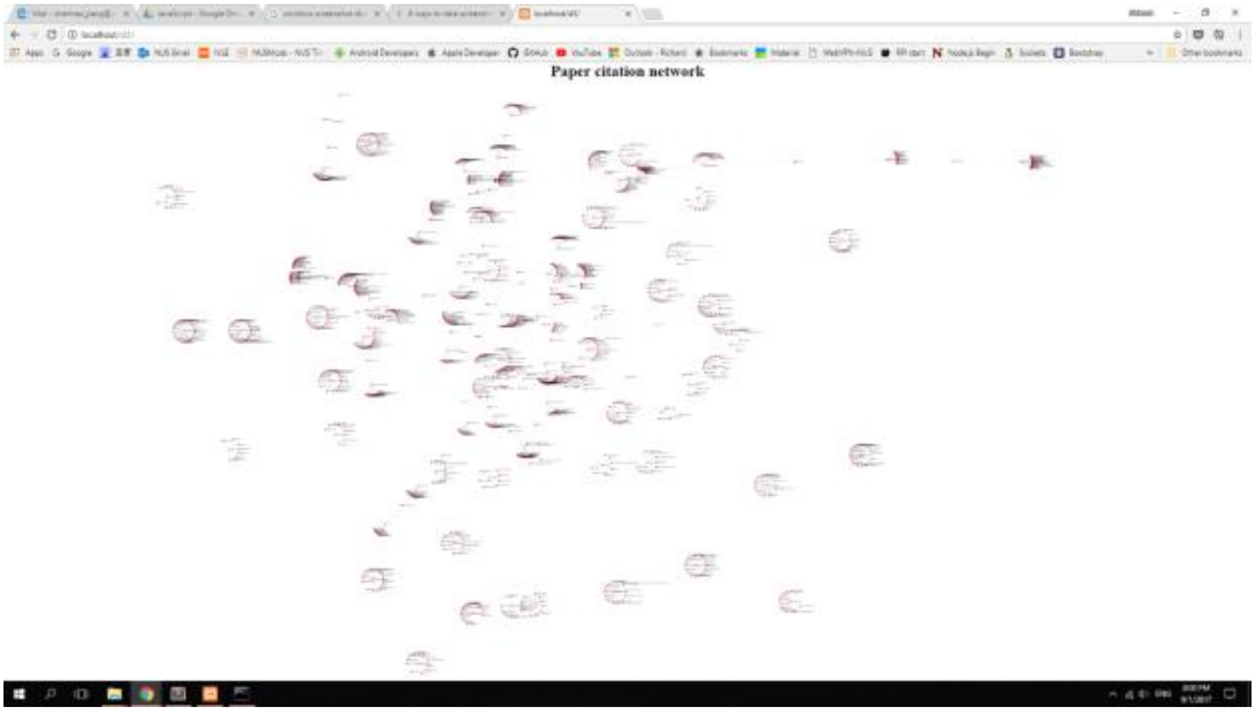
2. The top affiliations/universities:

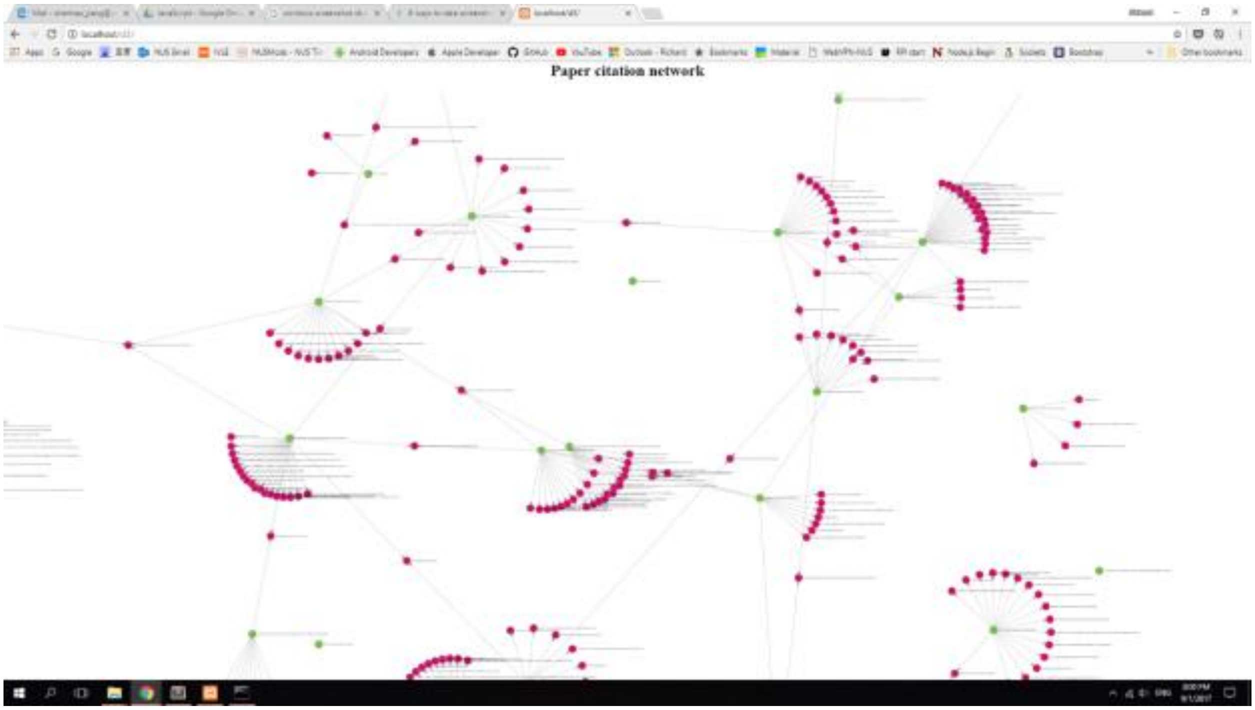


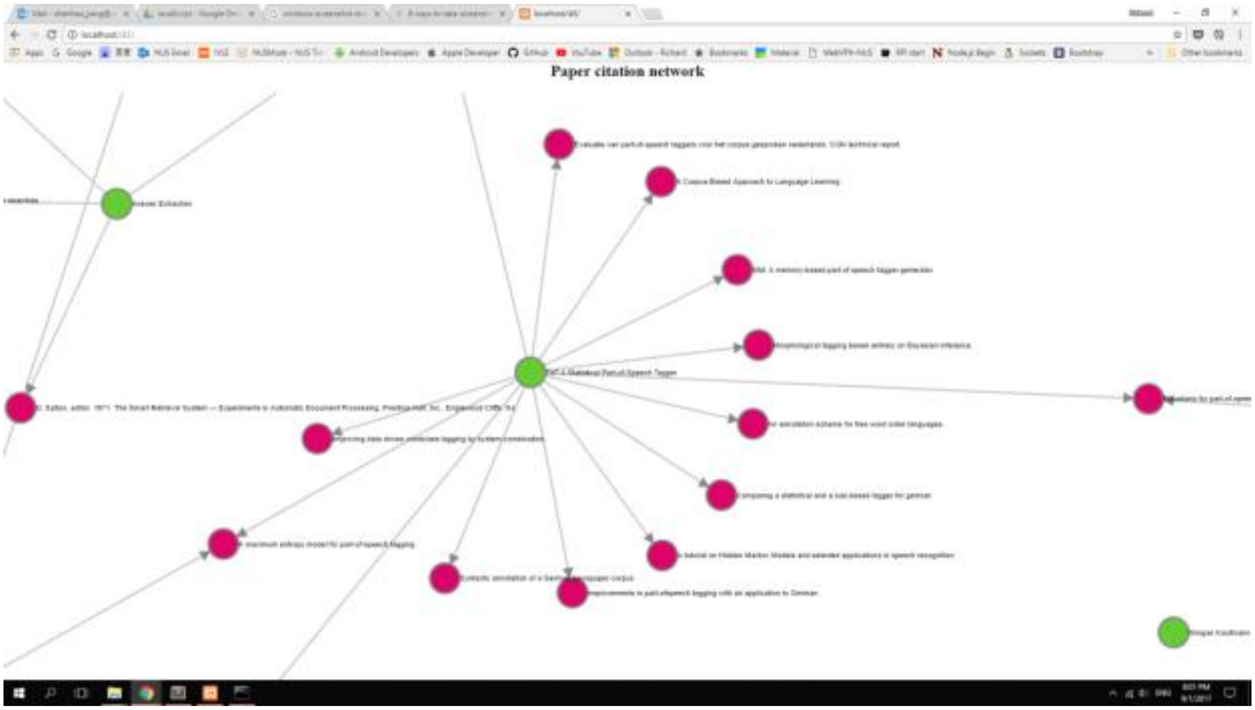
3. The trends of number of publications:



4. The citation web:

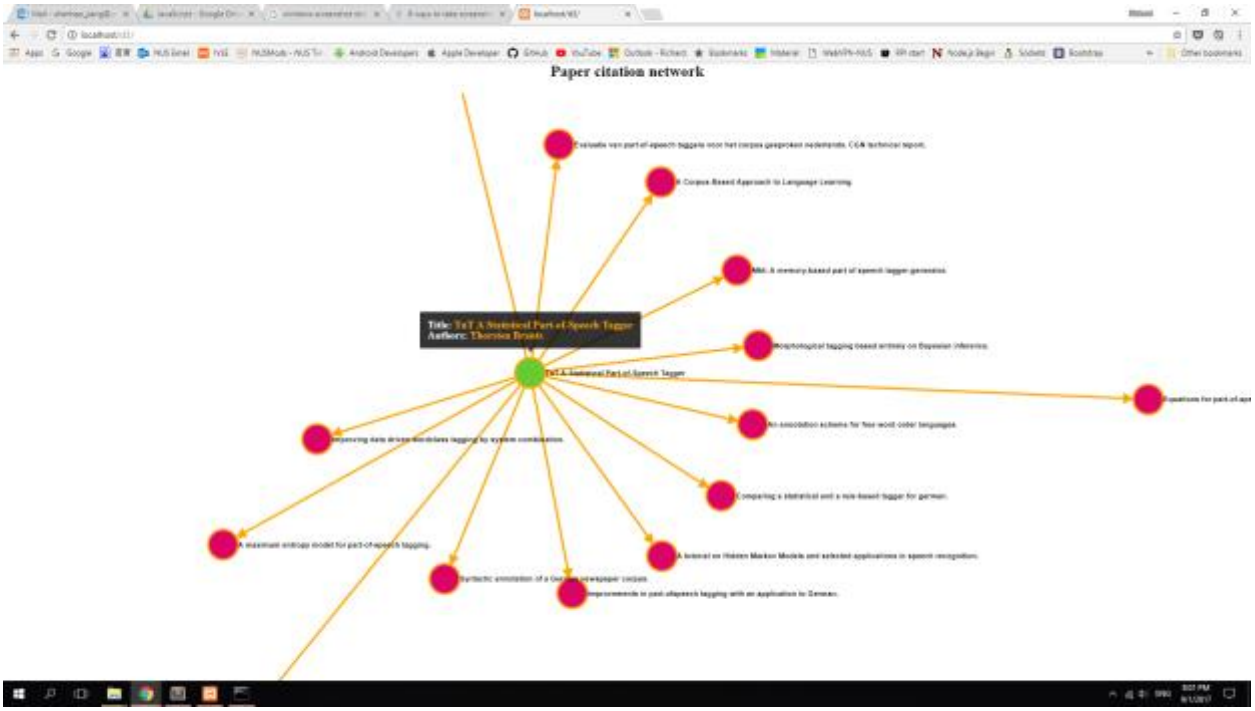






Paper citation network

Evaluatie van part-of-speech taggers voor het corpus gesproken nederlands. CGN technical report.
A Corpus-Based Approach to Language Learning.
Mbt: A memory-based part of speech tagger-generator.
Morphological tagging based entirely on Bayesian inference.
TnT-A Statistical Part-of-Speech Tagger
Equations for part-of-spee
G. Salton, editor. 1971. The Smart Retrieval System — Experiments in Automatic Document Processing. Prentice-Hall, Inc., Englewood Cliffs, NJ
Improving data driven wordclass tagging by system combination.
An annotation scheme for free word order languages.
Comparing a statistical and a rule-based tagger for german.
A maximum entropy model for part-of-speech tagging.
A tutorial on Hidden Markov Models and selected applications in speech recognition.
Syntactic annotation of a German newspaper corpus.
Improvements in part-ofspeech tagging with an application to German.



Paper citation network

Title: TnT A Statistical Part-of-Speech Tagger
Authors: Thorsten Brants

Evaluatie van part-of-speech taggers voor het corpus gesproken nederlands. CGN technical report.
A Corpus-Based Approach to Language Learning.
Mbt: A memory-based part of speech tagger-generator.
Morphological tagging based entirely on Bayesian inference.
TnT-A Statistical Part-of-Speech Tagger
Equations for part-of-spe
Improving data driven wordclass tagging by system combination.
An annotation scheme for free word order languages.
Comparing a statistical and a rule-based tagger for german.
A maximum entropy model for part-of-speech tagging.
A tutorial on Hidden Markov Models and selected applications in speech recognition.
Syntactic annotation of a German newspaper corpus.
Improvements in part-ofspeech tagging with an application to German.