Trier University Trier, 28.12.2022
Department IV – Computer Sciences
Prof. Dr.-Ing. Ralf Schenkel (H 539, `schenkel@uni-trier.de`)

Assignment 5 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, 05.12.2022

Note that the meetings on December 19, 2022 and January 2, 2023 will take place in the following Zoom meeting: `https://uni-trier.zoom.us/j/83637205106?pwd=RnowWVR6SUNJeHQ3MWZyeUtOV3Z3dz09` (Meeting-ID: **836 3720 5106**, password: **fHqDg15P**). It may be necessary to login with your university account before accessing the meeting.

Task 1: Thesauri 2+5+2+3+3 Points

In this task, we will use the (English) thesaurus WordNet: `http://wordnetweb.princeton.edu/perl/webwn`

(a) Determine in Wordnet all meanings of the noun **queen** and number them in an appropriate way.

(b) Which meaning of **queen** could me meant in the following sentences?

(1) In tarot decks, the queen outranks the knight.

(2) The Queen's 2012 Diamond Jubilee marked 60 years on the throne.

(3) Why Tina Turner is called "The Queen of Rock and Roll".

(4) Virgin queens go on mating flights away from their home colony to a drone congregation area.

(5) The Queen is the most powerful piece on the chessboard.

(6) "Queen" was the first album of the famous British rock band "Queen".

Which fundamental problem of thesauri can be observed here?

(c) Some knowledge bases rely on Wikipedia disambiguation pages for learning about different senses of a word. Which of the meanings from Wordnet can you identify in Wikipedia's disambiguation page for `Queen`, which is `https://en.wikipedia.org/wiki/Queen`?

(d) In practice, thesaurus entries are not assigned manually to term occurrences in a text, but automatically. A simple approach chooses for each term the most frequent meaning in the thesaurus. Explain why this is usually not a good approach.

(e) How could the assignment of term occurrences to thesaurus entries be done in a better way? Try to apply insights from author disambiguation.

Task 2: String Distances 2+4+9 Points

(a) Explore the **Hamming Distance** of two strings, for example in relevant literature or on the web. Briefly outline the essential principle of this distance. Specify your source(s).

(b) Explore the **Levenshtein Distance** and the **Damerau–Levenshtein Distance** of two strings, for example in relevant literature or on the web. Briefly outline the essential principles of these distances. Briefly explain how the two distances differ. Specify your source(s).

(c) Compute, if possible, the Hamming Distance, the Levenshtein Distance and the Damerau–Levenshtein Distance of **rhein** and **rhine**, of **mainz** and **minze**, and of **abababab** and **babababa**.

Task 3: Context-sensitive Spelling Correction 2+8+5 Points

Consider the word sequence
`A journey of thousand` *`meles`* `begins with a single step`
Obviously, the word `meles` is misspelled. Let possible correction terms be `miles`, `males` and `mules`. We now want to find the best among these correction terms in a context-sentive way with the bigram method.

(a) Specify which conditional probabilities you need to solve this task.

(b) Estimate these conditional probabilities with a search engine. For the conditional probability $P(x|y)$, you need to determine the number of results for the keyword **y** and the number of results for the phrase **„y x“**. Make sure that the search engine does not apply any automatic spelling correction. Which of the two candidates is the best correction term?

(c) Consider the query „zimk christmas“ (not meant as a phrase). What could be a possible information need behind this query? Use this example to explain possible problems and risks of an automatic spelling correction. Does a search engine propose a correction of this query (or even apply an automatic correction), and if so, which; try to explain the behavior of the search engine.

These tasks will be discussed in the tutorial on December 12.

General remarks:

- The tutorial group takes place on Mondays in the regular meeting in F55 at 12:15.
- The first meeting of the tutorial group will be on November 14, 2022.
- To be admitted to the final oral exam, you need to acquire at least 50% of the points in the assignments. In addition, you need to present at least one solution of a task in a convincing way during the tutorial.
- It is preferred to submit in groups of size up to two (but not larger); in that case, only one submission is sufficient for the whole group. Write the names of all group members on your solutions.
- Solutions must be handed in before the deadline
  - in Moodle (`https://moodle.uni-trier.de/`, course `DL-IR-2022`) as as a PDF or, if submitting multiple files, as an archive (.zip or comparable).

  Submissions that arrive after the deadline will not be considered. The name of at least one group member should occur in the file name of your submission. If you want to modify a previously uploaded solution, just re-upload your solution.
- Graded versions of your submissions will be returned in Moodle until the following tutorial.
- Announcements regarding the lecture **and** the tutorial group will be done in the StudIP course for the lecture.