SURV703 – Content Analysis
Essay 1
Elisabeth Linek

Task: Find recent and interesting examples of applied text analysis in your favorized subject area. Summarize them in a brief essay of about 500 words (+/- 10%). For this assignment, you are free to choose one out of two distinct tracks: (1) the industrial track or (2) the academic track.

The tracks

- The industrial track: Identify one or several examples of products, brands, or entire companies which use text analysis in an innovative way to tackle an interesting task. Please refrain from using generic and generally well-known examples such as Alexa or Siri. Appropriate sources include (but are not limited to) News and company web sites.
- The academic track: Select one or several related research papers that catch your interest. Provide a brief literature review summarizing the approached task, the proposed methodology and the most significant results. Please use different literature than the one provided in the Syllabus; feel free to leverage it as basis for a snowball search, though. You are not expected to understand the applied methods in detail (since you will yet have to learn about them in the upcoming lessons!) and mistakes based on unfamiliarity with the methods will not affect your grade.

General remarks
You are free to discuss one example in greater detail or to compare several examples in a concise manner. Cite sources whenever appropriate, but please adhere consistently to one citation style (of your choice) such as APA or numbered citations. Grading will be based on whether you display that you are able to discuss your chosen examples in a clear and comprehensive manner. Formal subjective criteria (style, layout) will not be decisive factors.

Thinking about NLP driven business models or industrial ideas, all types of search engines coming to mind instantly, as one of the most advanced text mining benefits used by billions of people every day. Ever since the human mankind has been able to write, text was produced, saved, stored and searched—an enormous amount of text, providing an enormous amount of data.

In order to pick one of the most striking business ideas, that embedded text mining within its core remit, I decided to shortly present a useful tool, that recently crossed my way in the context of the IPSDS program. The tool I am referring to is called *turn it in*[1], a so called online service, that is aiming to prevent plagiarism.

The basic idea of *turn it in* is, to compare submitted manuscripts, articles, papers or other types of texts with already published text documents, claimed by different authors. Based on a comparison, the program provides a scale, that mirrors the amount of text passages that are referring to other authors work, published before—an identification of possible plagiarism, that should be clarified. Furthermore, *turn it in* provides feedback regarding the amount of text that was found from other sources at first sight, adding a flag to the submitted document, that provides feedback using shades of traffic light colours.

To give an example, a document with less than 10 percent of "already published text" is classified as almost individual work. In terms of plagiarism it is categorized to be OK and marked with a green flag. In a more detailed view, the results of the comparison can be reviewed in a way, where *turn it in* marks passages that equal one-to-one other documents red and provides a direct link to the source of he original text, naming its author and background of its publication. *turn it in* describes its mission as "identify unoriginal content with the world's most effective plagiarism detection solution. Manage potential academic misconduct by highlighting similarities to the world's largest collection of internet, academic, and student paper content."[2]

In the context of the academical field, where hundreds of texts are produced, published and building a base for grades and academic degrees such a tool is very helpful! To check for plagiarism is a task, that would be

[1] The webpage of the program is the following: https://api.turnitin.com/

[2] Citation from website, https://api.turnitin.com/

very cumbersome and complex if it is done by hand. Its a time consuming task, whereas *turn it in* provides feedback within minutes.

Such a generous praise does almost instantly prepare the transition to a more critical review. In despite of the support, that such a useful tool provides, on a critical perspective I would name the aspect, that the amount of sources that are included in such a comparison task must be enormous. And the question derives, if it might be approximately broad or large enough to be on the save side with the result?

When is a sentence classified to be copied from somewhere else? Would it be enough to change some words by synonyms in order to be out of range for such a “plagiarism-detector”? What if the original is published with a delay so that original and a plagiarism are confounded with each other? Just to name some aspects, that are opening the stage for further questions.

The results of such a comparison should be validated in any case, marks, tags or flags may just give a first hint, they can not be seen as a reliable result on their own. The problem of so called false negatives is crossing my mind here in that context, which could lead to an almost philosophical discussion in terms of text mining abilities and its limitations.

I would conclude that such a tool, as the one that I picked to present here, can only be as smart as the person who validates the so prepared results.

Another quite impressive idea based on NLP algorithms or technologies is the Ngram viewer[3] maintained by google. It visualizes history or better said temporal progressions in an impressive way using text mining in terms of frequencies, which are directly transformed into diagrams. With that, it conveys a message based on words without words – a phenomenon! But I would leave track here, if I would now open up another interesting story...

[3] https://books.google.com/ngrams