WIP Profile Projects
Web Information Processing

Anne Schuth
Maarten de Rijke
WIP Profile Projects

• 4 weeks
• 6 credits
• 1 or two people per project
• An existing idea, or your own idea?
• Individual supervision
• Grade based on report (+code)
5 ideas

1. Characterizing local search
2. Names in novels
3. News article orientation
4. Summarizing topics over time
5. Entity normalization in news and web documents

• Your own idea?
1. Characterizing local search

• What?
  – Local search is an often used search application in which users have a very specific type of information need: users want to find and locate their place of interest (POI) on a map, often so that they can navigate to it. Queries consist of the name or the type of the POI the user is looking for plus a location. Typically, a user is not after more information about their POI, other than the location. When presented with a ranking of results for a query, user do not only click on these results, they interact with it in several ways: users call the POI, plan routes, actually drive to a POI, etc.

• Aim
  – Characterize local search logs from a major commercial local search engine. Among the questions that should be answered, are the following. Can we quantify bias and noise in user actions? How strong is, for instance, the bias introduced by the ordering of search results? In how far does the bias or noise depend on the type of device or query? How long are sessions and, related, how persistent are users? Do they reformulate queries until they find what they were looking for? When we assume that one interaction signal is ground truth, what’s the agreement with other signals? E.g., how often is a click on a result followed by a navigate-to action? Can a navigate-to action be used as ground truth?

• Main deliverable
  – A report that describes characteristics of the search logs

• Planning
  – First meeting at 16:00 today

• Who
  – Anne Schuth (me, A.G.Schuth@uva.nl)
2. Names in novels

• What?
  – How easy is it to read a translated novel? Professor of computational literature Karina van Dalen investigates this by looking at the names (of people, places, products, etc) occurring in novels. We have extracted names in 7000 Dutch novels and linked them to Wikipedia. These links allow us to say something about the "internationality" of a novel (e.g. how many languages have a Wikipedia page about a name in a novel).

• Aim
  – A good measure for the task, evaluation of its robustness, scores on a 7000 book collection

• Deliverable
  – Write-up + code

• Planning
  – First meeting tomorrow morning

• Who
  – Maarten Marx (M.J.Marx@uva.nl)
3. News article orientation

• What
  – When trying to determine the overall sentiment of users (comments) towards a topic we often start with collecting news articles about that topic. To correctly interpret the comments it is necessary to understand the orientation of the news article towards the topic. For example, comments on an article that says "Joint Strike Fighter outperforms Eurofighter in test" should be interpreted differently than those on one that says "Budget cuts lead to purchase of less JSFs".

• Aim
  – define the problem of news article orientation, explore literature to find possible solutions, and perform a set of experiments to see if this task can be solved.

• Deliverable
  – Write-up + code

• Planning
  – First meeting tomorrow morning

• Who?
  – Wouter Weerkamp (w.weerkamp@uva.nl)
4. Summarizing topics over time

- **What**
  - News topics that exist for a longer period of time tend to shift in their focus. Again, looking at the JSF, initially the discussion was about the profit Dutch companies would make on producing it, whereas now the discussion is on whether we should still buy the planes. Can we, using news articles and their comments, generate a summary that shows how the focus of the topic changed over time?

- **Aim**
  - Explore the existing summarization literature and implement one or more methods. An evaluation methodology should be proposed that can be used to test the effectiveness of the system.

- **Deliverable**
  - Write-up + code

- **Planning**
  - First meeting tomorrow morning

- **Who**
  - **Wouter Weerkamp** (w.weerkamp@uva.nl)
5. Entity normalization in news and web documents

• What
  – The task of entity normalization is to map name strings (entity mentions) to their unambiguous referents (real world entities). Given a dataset consisting of document-entity mention pairs, the task is to cluster the mentions that refer to the same entity. The challenge is ambiguity: identical mentions can refer to different entities, and different mentions can refer to a single entity.

• Aim
  – Contribute to UvA TAC submission

• Deliverable
  – Write-up + code (module that, given a list of queries (query ID + document ID), clusters the associated entity mentions and returns a list of query ID + entity ID pairs)

• Planning
  – First meeting at 16:00 today

• Who
  – David Graus (D.P.Graus@uva.nl)