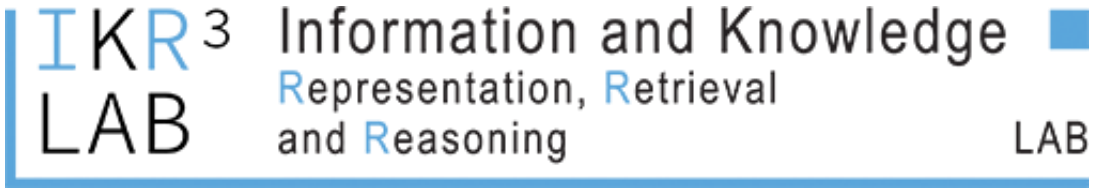
Information Retrieval

2019/2020 Projects



General requirements

- The project must be done in groups of 3 people **at most**.
- **Requirements**:
  - Delivery of **all the material** necessary to install and run the developed system:
    - A **README.txt** document of the how-to install and run the system;
    - **Source code**;
    - **Binaries** and necessary **libraries**.
  - A (detailed) **report** describing the system and the choices undertaken.
  - A **PowerPoint presentation** illustrating the system and the choices undertaken. There will be an oral presentation and a discussion.
  - The report and the PowerPoint presentation must be written **in English**.

General requirements

- The implementation of **more advanced Lucene techniques** will be positively evaluated.
- **Additional functionalities** at your choice can be implemented.
  - They will be positively evaluated.
- The documentation and the source code of the system must be sent to the Professor and the teaching assistant at least **7 days BEFORE the written exam**.

IKR³ LAB

A personalized search engine for microblog contents (but not only…)

PROJECT A

Project A – Goal

- Create a **personalized search engine for tweets** (e.g., news) or other **textual contents** based on the following requirements:
  - Create **multi-layer user profiles** for 10 users representing their interests (user profiles can be represented as bag-of-words).
  - Provide each target user with a **different personalized ranked list** of tweets (or other textual contents) in response to a query **and the user interests**.
  - In addition to **topicality**, the system should combine one or more **additional relevance dimensions** to rank tweets or other textual contents.

Project A – Dataset

- **Two possibilities**:
  - Crawl specific tweets by using the Twitter APIs (at least for one month) for **5 categories of content** (e.g., news categories: sport, cinema, music, …).
    - In the case of news, it is possible to crawl Twitter news by **Twitter news accounts**.
  - Find suitable **textual large-scale datasets** from which extract suitable textual contents to be retrieved, belonging to **5 categories of content**.
    - In this case, the **size** of the dataset, its **adequacy** to the considered problem and the **pre-processing phase** will be evaluated.

Project A – User profiles

- Different “layers” of the profile represent different interests, to which different keywords are associated.

- The keywords associated with each interest will be **automatically extracted from some textual documents that can represent the user’s interests**.

IKR³ LAB

A Recommender System for microblog contents (but not only…)

PROJECT B

Project B – Goal

- Create a **Recommender System for tweets** (e.g., news) or other **textual contents** based on the following requirements:
  - Create **multi-layer user profiles** for 10 users representing their interests (user profiles can be represented as bag-of-words).
  - Provide each target user with **different recommendations** of tweets (or textual contents) **based on her/his interests**.

Project B – Dataset

- **Two possibilities**:
  - Crawl specific tweets by using the Twitter APIs (at least for one month) for **5 categories of content** (e.g., news categories: sport, celebrities, cinema, music, etc.).
    - In the case of news, it is possible to crawl Twitter news by **Twitter news accounts**.
  - Find suitable **textual large-scale datasets** from which extract suitable textual contents to be recommended, belonging to **5 categories of content**.
    - In this case, the **size** of the dataset, its **adequacy** to the considered problem and the **pre-processing phase** will be evaluated.

Project B – User profiles

- Different “layers” of the profile represent different interests, to which different keywords are associated.

- The keywords associated with each interest will be **automatically extracted from some textual documents that can represent the user’s interests**.

Project B – Additional information

- It is possible to provide recommendations **separately** for each interest, or **grouped together** with respect to topical interests of the user.

- It is possible to **let the user specify** for which topical interest/s s/he wants receive recommendations.

Useful links

- **Twitter4J** unofficial Java library for Twitter API.
  - To be downloaded at the following link:
    - http://twitter4j.org/en/index.html

- **Twitter APIs:**
  - To be downloaded at the following link:
    - https://developer.twitter.com/en/docs