Wikipedia recommender system

Deadline 16 days after the 4th lab, send report to my email, start with [IR]
grzegorz.miebs@cs.put.poznan.pl

Can be done in pairs

The general task is to create a system that will recommend similar articles based on the previously visited articles.

Input - Collection of articles (links or titles)
Output - Collection of recommended articles (links or titles) with a "score"

You will receive a grade for each of the following steps. The highest possible score without finishing all parts is 4.0. For example, if you do perfectly the first two steps your grades will be 4.0, 4.0, 2.0.
Crawling and scraping - Download text from at least 1000 Wikipedia/fandom wiki articles. (Scrappy is not a must)
Stemming, lemmatization - preprocess downloaded documents into the most suitable form for this task. Store it as a .csv/parquet file or into a database.
Similarities - for a given collection of previously visited articles find the best matches in your database and recommend them to the user

GUI not required, notebook or any other reasonable form will be accepted. I have to be able to provide a list of articles in an easy way and receive a meaningful recommendation.
You have to send the source code and report.

Report:
- pdf or notebook
- explain each step of your algorithm, especially how you score articles
- present interesting statistics about your database (most frequent words, histograms, similarities between documents, ...)
- show some examples of recommendations with explanations (I'd prefer graphical form - see prediction breakdowns for example)