Lab1

Ludvig Noring - ludno249
Michael Sörsäter - micso554

Implementation

Web Crawling

From the Google play store page we extract all links to the apps. From those apps we extract apps until we have enough apps.

From each app we filter out the description and the title.

Create Inverted Index

The description for the app is tokenized, stemmed, stopwords are removed and non-ascii characters are removed.

Each term in the vocabulary is mapped to an index.

Idf is calculated for each term by taking the logarithm of the number of apps divided by the number of documents the term occurs in.

The normalized tf is calculated for each term and document by taking the count of the term in the document divided by the count of the most frequent term in the document.

Each document have a vector (with the same length as the vocabulary). For each term the value is calculated by multiplying tf and idf.

Calculate similarity

The query is processed the same as the documents, resulting in a vector.
This vector is compared with the vectors for each app with the following formula:

$$sim(vector_{app}, vector_{query}) = \frac{vector_{app} \cdot vector_{query}}{|vector_{app}||vector_{query}|}$$

The k best apps are returned.

Test runs

>>> Enter query and k: my phone is slow 5
Turbo Cleaner - Boost, Clean
Power Clean - Anti Virus Cleaner and Booster App
Speed Booster - Ram, Battery & Game Speed Booster
Antivirus & Mobile Security
CShare (File Transfer Tools)

>>> Enter query and k: edit photos crop 5
Google Photos
Flickr
Google Docs
Candy Gallery -Photo Edit,Video Editor,Pic Collage
BeautyPlus - Easy Photo Editor

>>> Enter query and k: dead boys walking 5
The walking zombie: Dead city
The Walking Dead No Man's Land
Stick Hero
The Walking Dead Dead Yourself
Noom Walk Pedometer

>>> Enter query and k: card game alone 5
Shuffle Cats
Spider Solitaire
Solitaire Free
Solitaire card game
Solitaire