Predicting news story keywords

Jan Viilma

April 3, 2018

Contents

From the proposed project topics it is not clearly stated wether the prediction is done by only using the headline of a news story or the content of the story. From the given example:

- "Manchesteri suurklubid võivad rüseluse eest karistada saada" -> manchester city, manchester united, jalgpall

it could be derived that only the headline is used for predictions. In either case the project involves topic modelling, which is a natural language processing method used for discovering the abstract keywords that occur in a text. For a comprehensive topic extraction gensim could be used. If only the headline is used, my first approach would be to extract the nouns of that headline and implement a python NLP library that finds words with similar meaning. For further improvement gensim could be used again to not only extract nouns but also phrases or collocations - words which most likely go together. Gensim uses the most popular topic modelling algorithms, such as Latent Semantic Analysis, Latent Dirichlet Allocation and Random Projections.