[ptt1819] Assignment 4
Topic Extraction

A. K. and A. S.

February 9, 2019

Approach and Expectations

- Question:
- Hypothesis:

Overview

1. WikiDump(pages& categorylinks) - import in mysql - retrieve categories and put them in a .csv
2. extract relevant categories and their page id (python script - json) - split wiki articles multistream?
3. LDA (scala + spark + MLlib) - preprocessing - results in graph/chart-form (pandas?, for spark - Vegas/breeze)

Collect Data

Topic Modeling with LDA - Idea

-

Topic Modeling with LDA - Implementation

Preprocessing

Preliminary Results

Q&A

Questions?