Big data project

In this project we aim to group Vietnamese words into topics. Each topic will consist of words with related meanings. For example there could be a topic with sport-related words such as football, Olympics, Wimbledon, Messi…. The topics are created automatically by clustering word vectors, and we will generate these word vectors from the text in the Vietnamese language Wikipedia.

The following steps are involved:

1. Download Vietnamese wiki from wikidump: [viwiki-20150702-pages-meta-current.xml.bz2](https://dumps.wikimedia.org/viwiki/20150702/viwiki-20150702-pages-meta-current.xml.bz2). (Extract this and use <https://www.emeditor.com/download/> to open).

2. Remove wiki code (headers, tables, links, picture tags…).

3. Convert XML escape characters to UTF-8.

4. Tokenize the text with a Vietnamese tokenizer application (using <http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer> )

5. Generate word vectors from the processed text with the word2vec library (using <https://code.google.com/p/word2vec/> or any other tools support this one).

6. Cluster the resulting word vectors in to topics (using Mahout to do, you can read chapter 7,8 and 9.1 in “Mahout in action” ebook, download from <http://openresearch.baidu.com/u/cms/www/201210/30144944cqmu.pdf;jsessionid=AD8D0E0752C49DAF177F094ECB01FDEE> ).

Upon completion of the project, you will be able to:

1. Process XML and plain text using regular expressions and hadoop.

2. Use the word2vec library to generate word vectors from preprocessed text.

3. Perform a standard clustering algorithm on a large dataset.

You have to send me your source code and report about the results (in zip file named *group\_name.zip*) before ***7 Aug 2015 12PM***.