Theme 3.4 System Extension.

Build Mongolian language package compatible with the Spacy NLP toolkit as alpha support model

Goal: Build Mongolian language package compatible with Spacy NLP toolkit as alpha support. Create it as open-source project. Try to contribute and integrate it with the Spacy toolkit extensions list.
Research: Spacy toolkit extensions. Look and build the stop-words dataset for Mongolian language. Build some basic lexical attributes, punctuations using regex, build some basic tokenization exceptions for Mongolian language.

Requirements

1. Team: Single person because it's related to a specific foreign language - Mongolian language, not sure any other student knows Mongolian language
    1. Name: Bayar Demberel
    2. NetID: bd12@illinois.edu
    3. Captain: Bayar Demberel
2. Chosen system: Spacy NLP toolkit.
    1. Subtopic: Adding models for new languages
3. Build the stop words dataset for Mongolian language, build some basic lexical attributes, punctuations using regex, build some basic tokenization exceptions.
4. Once built the Mongolian language extension, people can use it with the Spacy toolkit to process texts on Mongolian language. Initial model will be able to tokenize given text with Spacy toolkit. The Spacy toolkit does not have included the Mongolian language yet, so I excited to decide to work on this.
5. Once if it successfully contributed to Spacy extensions, any developers could use it with Spacy toolkit for NLP for Mongolian language. If it could not yet able to contribute it to Spacy community, it will be still showing as publicly and developers may can download and use this package.
6. Language to use: Python
7. Total estimated hours for this extension would be at least 40 hours. I am really wanted to do this, so I will spend as much time as I can.
    1. Research about Spacy
    2. Build stop-words dataset for Mongolian language – May need to build some web crawlers to collect large amount of text and create the stop-words list.
    3. Build some basic lexical attributes for Mongolian language
    4. Build some punctuations
    5. Build some basic word tokenization exceptions for Mongolian language

The whole project will be publicly contributed as open-source community, so any developers can improve the system, it is open to anybody wanted to contribute.