Progress report

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

1. Research about Spacy
2. Build stop-words dataset for Mongolian language
3. Build some basic lexical attributes for Mongolian language
4. Build some punctuations
5. Build some basic word tokenization exceptions for Mongolian language

Completed

- Forked the spaCy repo from github
- Added the folder and specific required files for Mongolian language package into my repo
- Created some example sentences for Mongolian language to test spaCy and its language model
- Created the lex_attr.py file which can detect numerical words for Mongolian language
- Created punctuation.py
- Added initial version of Mongolian stop-words list using an existing github source
- Created tokenizer_exceptions.py which created some initial versions of abbreviations for Mongolian language, for example weekdays abbr. etc.

Pending

- Compile the spaCy repo
- Create test cases for Mongolian language model
- Improve the stop words
- Improve the abbreviations and punctuations
- Test changes
- Create the running version for Mongolian language model compatible with spaCy
- Commit to the spaCy community
- spaCy website changes if the change published

Challenges

- Compiling the whole spaCy repo
- Mongolian stop words
- Abbreviations and exceptions words in Mongolian languages
- Test changes
- Push to spaCy
- Running version of my model