Joshua Singh, Cullen Paulisick,
Ryan Schneider, Nikhil Gupta

SaturnSearch

Our goal was to create a search engine which researchers can use to find existing genetic parts for use in synthetic design based on user-specified abstract criteria (function, use, system, etc). Upon finding the desired genetic part, the user can export the Json file associated with the part along with its Genbank sequence for further downstream genetic design. This will reduce the time it takes for researchers to find genetic parts during the specification/design phase and allow for higher level design abstraction. To begin, the user is prompted to enter a search phrase which is separated into keywords and fed into an unsupervised LDA model. This model creates a topic distribution from the user's query phrase and then matches these topics to relevant papers. The relevant papers are subsequently mined for genetic parts using *IBM-Watson Natural Language Understanding* and the genetic parts from the most relevant papers are identified. From this part list, the user can click on a specific part and the relevant DNA sequence from GenBank is displayed. Therefore, by using *SaturnSearch*, researchers can avoid time-consuming manual literature searches, find relevant genetic parts and store their pertinent information in easy to use data structures for further downstream design.

Software Components:

Gensim:

Open-source library for unsupervised topic modeling and natural language processing. Gensim was used to import our LDA model framework and implement a "bag-of-words" document representation for the papers obtained from the PMC database.

NCBI E-utilities API:

Provides an interface to query the NCBI database system. For our project, only articles from the PMC database were used as they are all open-source and full-text. The relevant texts were queried and downloaded based on their PMCIDs.

NLTK:

Natural language toolkit for python that allowed for language processing operations such as tokenizing, lemmatizing and creating bigrams/trigrams. This allowed us to process the raw text documents into a more useful format for term extraction.

Biopython:

Open-source biological computation tool used for its *Entrez* parser module which provides useful functions for querying the PMC database. Specifically, this module allowed us to search for PMC papers relevant to the user-specified query terms and store these papers as Json structures.

PyQt:

Python binding of cross-platform GUI toolkit which was used to develop our GUI. PyQt was chosen due to its ability to produce high quality, modern-looking and easy to use GUIs. Additionally, this GUI toolkit was chosen due to its ability to easily integrate with our main python scripts.

IBM-Watson Natural Language Understanding:

An open-source deep learning product which can be used to extract entities from text documents. The papers deemed relevant to the user query search are fed into this software and the deep learning model extracts genetic parts. Previous to this extraction, the model is trained to recognize genetic part names through user annotation.

Compiling Instructions:

Project Setup:

1. Create a Python Virtual Environment
2. Use pip to download dependencies from the dependencies.txt : i.e python3 -m pip install -r dependencies.txt
3. Run the command : python SaturnSearchAlpha.py

IBM Watson API: The API key for the Watson has to be dynamically changed so as to avoid security issues from being raised. The login credentials for this application will be sent to graders via email so as to not compromise them. Examples of our training data and model can also be found using these credentials at https://us-east.knowledge-studio.watson.cloud.ibm.com/knowledge-studio/tools/app/rzhz10/thb30lw84b4tl27s/ui/#/. Deployed model ID will also be provided.

ISSUE RAISED: A ticket has currently been opened with IBM due to an issue with accessing the API. This may be due to our API key being compromised. Updates will be posted to the github.

Dependencies: The dependencies are as follows and can be installed/accessed from our "dependencies.txt" file on the github.

---

biopython 1.78 certifi 2020.12.5 chardet 4.0.0 click 7.1.2 Cython 0.29.21 gensim 4.0.1 ibm-cloud-sdk-core 3.9.0 ibm-watson 5.1.0 idna 2.10 joblib 1.0.1 json-flatten 0.1 nltk 3.6.2 numpy 1.20.2 pip 21.0.1 PyJWT 2.0.1 PyQt5 5.15.4 PyQt5-Qt5 5.15.2 PyQt5-sip 12.8.1 python-dateutil 2.8.1 python-Levenshtein 0.12.2 regex 2021.4.4 requests 2.25.1 scipy 1.6.2 setuptools 56.0.0 six 1.15.0 smart-open 5.0.0 tqdm 4.60.0 urllib3 1.26.4 websocket-client 0.48.0 wheel 0.36.2 xmltodict 0.12.0