Python Frameworks

	NLTK
Open Source/Closed Source	• Easy platform to work with human Language data • Open source and provides 50 corporas and lexical resources • Many third party extensions
Framework/ Library	• Library with plenty of approaches to each task • Fast sentence tokenization • No support for semantic structure • No support for Neural network models
Integration APIs	Support for various API references , listed below:- • Collocation module • Data module • Downloader module • featStruct module • Grammar module • Help module • Probability module • Text module • Toolbox module • Tree module • Util module Check full documentation here - https://www.nltk.org/api/nltk.html
Language Compatibility	• By downloading *Punkt* several langauges are compatible with NLTK , Czech, Danish, dutch, English, Estonian, finnish, French, german, greek, Italian, Norwegian, polish, Portuguese, Slovene, Spanish, Swedish, Turkish
Number of Years	Active from April 2011 , first version 2.0.1rc1 (latest version 3.4.5)

	GENSIM
Open Source/Closed Source	• Gensim is an Open source Python library available under GNU license v2 for topic modelling, document indexing and similarity retrieval with large corpora • Corpora is plural for Corpus (a bag of text phrases in form of sentence, paragraphs) • Depends on two peer libraries used by Python – NumPy and SciPy.
Framework / Library	• Gensim is library which is extremely useful NLP library primarily developed for TOPIC MODELLING. • However , it supports a variety of other NLP tasks such as converting words to vector, documents to vector, finding text similarity and text Summarization. (Topic modelling is used to Extract hidden topics from large volume of text. LDA algorithm is used here. However extracting good quality of topics that are clear, segregated and meaningful

	depends on quality of text preprocessing and strategy of finding optimal number of topics)
Integration APIs	Gensim provides an inbuilt API to download popular text datasets and word embedding models.A comprehensive list of available datasets and models is maintained here - https://raw.githubusercontent.com/RaRe-Technologies/gensim-data/master/list.json Using the API to download the dataset is as simple as calling the *api.load()* method with the right data or model name.
Language Support	
Number of Years	The first version of Gensim 0.2 was introduced in 2010. The latest version which is currently active is 3.8.1 with support for Python 3 and successors.

	CORE NLP
Open Source/ Closed Source	Much of this software can easily be used from Python (or Jython), Ruby, Perl, Javascript, F#, and other .NET and JVM languages. These software distributions are *open source*, licensed under the GNU General Public License (v3 or later for Stanford CoreNLP **(Although many features are not supported by NLP standards which makes it less preferable for our project)**
Framework/ Library	Stanford Core NLP is a not a framework , it is an Integrated toolkit with broad range of grammatical analysis tools. Used in integration with various Stanford NLP tools such as POS(Part of Speech tagger), Named Entity Recognizer(NER), Sentiment Analysis, etc.
Integration APIs	***Simple CoreNLP API*** Simple API for users who wants to avoid excess customization. Supported Annotator for this API reference are- • Tokenization • Sentence Splitting • Part of Speech Tagging] • Lemmatization • Named Entity Recognization • Dependency Parsing • Conference Resolution • Natural Logic Polarity Miscellaneous Extra provides support for *Sentence Algorithms.* Refer below link to get an overview about how to use CoreNLP API in our project: https://stanfordnlp.github.io/CoreNLP/api.html
Language Support	Latest release has support for following languages- • Arabic • Chinese • English

	• English [KBP] (Due to size issues , English language has been split into two jars. KBP contains extra resources needed to run relation extraction and entity linking) • French • German • Spanish
Number of Years	9 years Initial Release – January 2010 [version 1.0] Latest Release – October 2018 [version 3.9.2]

	SpaCy
Open Source/ Closed Source	It is an Open Source Library for advanced NLP written in C and Python. Allows for Cross platform operation
Library	The library provides extensive features as listed- • Non-destructive Tokenization • Named Entity Recognition (NER) • Pretrained vectors • POS tagging • Syntax driven Sentence Segmentation • Binary serialization
Integration API	Central data structures in SpaCy are – *doc* and *vocab.* Check the below link for Complete API reference:- https://spacy.io/api
Language Support	SpaCy provides support for 53 languages including common ones and many new ones. Some of the interesting new languages are listed below:- Greek, Lithuanian, Afrikaans, Bengali, Catalan, Telugu, Estonian, Kannada, Marathi, Latvian, Sinhala, Tagalog, Tatar, Urdu. Some of the language tokenizers require external dependencies- **Russian –** https://github.com/kmike/pymorphy2 **Ukranian –** https://github.com/kmike/pymorphy2 **Thai –** https://github.com/wannaphongcom/pythainlp **Chinese –** https://github.com/fxsjy/jieba **Japanese –** http://unidic.ninjal.ac.jp/back_number#unidic_cwj **Korean –** https://bitbucket.org/eunjeon/mecab-ko/src/master/README.md **Vietnamese –** https://github.com/trungtv/pyvi
Number of years	4 years [Initial release Febrary 2015]