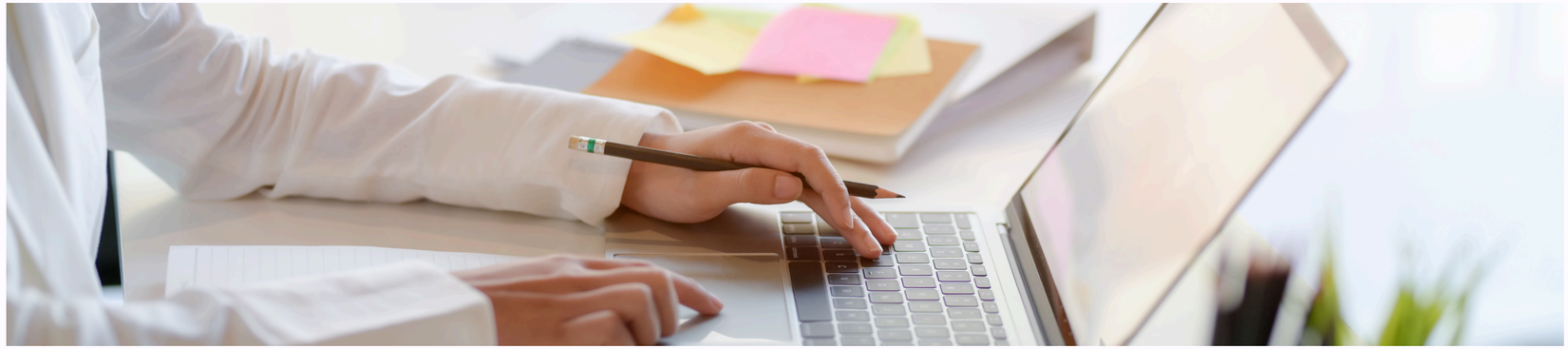
Жасанды интеллект

PYTHON ҮШІН 8 ҮЗДІК ТАБИҒИ ТІЛДІ ӨҢДЕУ (NLP) КІТАПХАНАСЫ

Ниязбекова Жұлдыз

Тарих-география 301

NLP ДЕГЕН НЕ?



Natural Language Processing (NLP) – бұл компьютерлерге адамның табиғи тілін (мысалы, қазақ, ағылшын, орыс және басқа тілдер) түсінуге, өңдеуге және генерациялауға мүмкіндік беретін жасанды интеллект (AI) саласы.

NLP НЕГІЗГІ МАҚСАТТАРЫ:

01

Табиғи тілді түсіну (NLU – Natural Language Understanding): мәтіннің немесе сөйлеудің мағынасын талдау, сұрақтарға жауап беру, мәтінді санаттарға бөлу.

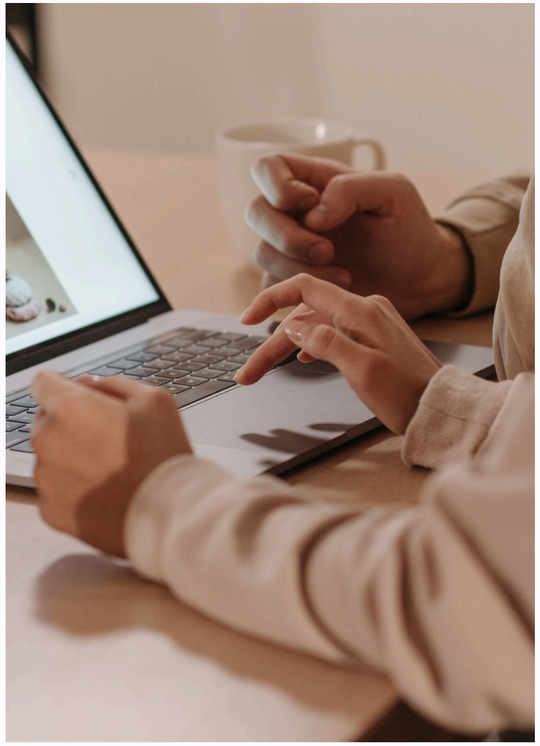
02

Табиғи тілді генерациялау (NLG – Natural Language Generation): адамға түсінікті мәтін құру (мысалы, жаңалықтарды автоматты түрде жазу, чат-боттармен сөйлесу).

03

Сөзді алдын ала өңдеу: морфологиялық, синтаксистік және семантикалық талдау.

NLP ҚОЛДАНЫЛУ САЛАЛАРЫ:



- Автоаударма (Google Translate, DeepL)
- Дауыстық көмекшілер (Siri, Alexa, Google Assistant)
- Чат-боттар (клиенттерге қызмет көрсету жүйелері)
- Спам сүзгілеу (электрондық поштадағы қажетсіз хаттарды анықтау)
- Сезімдерді талдау (пікірлерді "оң", "теріс", "бейтарап" деп анықтау)

NLTK

NLTK (Natural Language Toolkit) – бұл Python бағдарламалау тілінде жазылған табиғи тілді өңдеуге (NLP) арналған кітапхана. Ол мәтіндерді талдау, өңдеу және машиналық оқыту модельдерін қолдану үшін көптеген құралдар ұсынады.

Мүмкіндіктері

- Мәтінді токенизациялау (сөздер мен сөйлемдерді бөлу)
- Стемминг және лемматизация (сөздердің түбірін табу)
- Синтаксистік және морфологиялық талдау
- Құрамында алдын ала дайындалған корпус пен лексикалық ресурстар бар (мысалы, WordNet)
- N-грамма талдауы (сөз тіркестерін анықтау)
- Машиналық оқыту алгоритмдері (классификация, кластерлеу)

TEXTBLO

TextBlob – бұл Python тілінде табиғи тілді өңдеуге (NLP) арналған қарапайым және ыңғайлы кітапхана. Ол NLTKжәне Pattern кітапханаларының негізінде жасалған және мәтінді өңдеуге, талдауға, сондай-ақ машиналық оқыту әдістерін қолдануға мүмкіндік береді.

TextBlob негізгі мүмкіндіктері:

- Мәтінді токенизациялау (сөздер мен сөйлемдерді бөлу)
- Сөздерді лемматизациялау (сөздің бастапқы түрін анықтау)
- Синтаксистік және морфологиялық талдау
- Сезімдерді талдау (пікірдің оң немесе теріс екенін анықтау)
- Автоаударма және тіл анықтау
- N-грамма генерациясы (сөз тіркестерін анықтау)
- Орфографияны тексеру және түзету

TextBlob мен NLTK айырмашылығы

- TextBlob қолдануға жеңіл және қарапай ым, негізінен шағын NLP жобаларға арналған.
- NLTK көбірек функционалдылық ұсынады, бірақ оны пайдалану күрделірек

CoreNLP – бұл Stanford University жасаған табиғи тілді өңдеуге (NLP) арналған қуатты кітапхана. Ол Javaтілінде жазылған және көптеген NLP тапсырмаларын орындай алады. CoreNLP жоғары дәлдігімен ерекшеленеді және кең ауқымды лингвистикалық талдау құралдарын ұсынады.

CORENLP НЕГІЗГІ МҮМКІНДІКТЕРІ:

- Токенизация (мәтінді сөздер мен сөйлемдерге бөлу)
- POS-тегтеу (сөз таптарын анықтау – етістік, зат есім, сын есім және т.б.)
- Лемматизация (сөздің түбірін анықтау)
- Атаулы объектілерді тану (NER) (адам аттары, ұйымдар, жер атаулары және т.б.)
- Синтаксистік талдау (грамматикалық құрылымды анықтау)
- Сезімдерді талдау (мәтіннің позитивті, негативті немесе бейтарап екенін анықтау)
- Coreference Resolution (мәтіндегі есімдіктердің қай объектілерге сілтеме жасайтынын анықтау)

GENSIM

Gensim – бұл Python тілінде жазылған мәтіндік деректерді өңдеуге және тақырыптық модельдеуге арналған кітапхана. Ол үлкен мәтіндік корпустардан семантикалық ақпарат алуға және оларды статистикалық әдістер арқылы өңдеуге мүмкіндік береді.

МҮМКІНДІКТЕРІ

01

- Векторизация – мәтінді сандық форматқа айналдыру
- Тақырыптық модельдеу (LDA, LSI, HDP алгоритмдері)

02

- Word2Vec – сөздерді вектор түрінде көрсету және олардың мағыналық жақындығын анықтау
- Doc2Vec – құжаттар арасындағы семантикалық байланысты анықтау

03

- FastText – сирек кездесетін сөздермен де жұмыс істей алатын сөздік векторизациялау әдісі
- Мәтіндік деректерді индекстеу және іздеу

SPACY

spaCy – бұл Python және Cython тілдерінде жазылған жылдам әрі тиімді табиғи тілді өңдеу (NLP) кітапханасы. Ол үлкен көлемдегі мәтіндерді талдау мен өңдеуге оңтайландырылған және машиналық оқытуға негізделген.
spaCy негізгі мүмкіндіктері:

- Токенизация – мәтінді сөздер мен сөйлемдерге бөлу
- POS-тегтеу – сөз таптарын (етістік, зат есім, сын есім және т.б.) анықтау
- Лемматизация – сөздердің бастапқы түрін табу
- Атаулы объектілерді тану (NER) – адам аттары, ұйымдар, жер атаулары және т.б.
- Зависимдік синтаксистік талдау – сөйлемдегі сөздердің өзара байланысын анықтау
- Сөздер арасындағы ұқсастықты өлшеу – семантикалық жақындықты бағалау
- Қарқынды мәтіндік өңдеу – үлкен мәтіндермен жұмыс істеуге оңтайландырылған
- Глубокая интеграция с машинным обучением – TensorFlow және PyTorch секілді кітапханалармен жұмыс істей алады

POLYGLOT

Polyglot – бұл көптілді табиғи тілді өңдеуге (NLP) арналған Python кітапханасы. Ол бірнеше тілдегі мәтіндерді талдау және өңдеу мүмкіндігін ұсынады, сондықтан көптілді NLP жобаларында жиі қолданылады.

- 160+ тілде мәтін өңдеу
- Токенизация – сөздер мен сөйлемдерді бөлу
- Сөз таптарын анықтау (POS-тегтеу)
- Атаулы объектілерді тану (NER) – адам аттары, ұйымдар, жер атаулары және т.б.

- Сезімдерді талдау – мәтіннің позитивті, негативті немесе бейтарап екенін анықтау
- Мәтіннің тілін автоматты анықтау
- Сөздер арасындағы мағыналық байланыс пен векторизация

Scikit-learn – бұл Python тіліндегі машиналық оқытуға (ML) арналған ең танымал кітапханалардың бірі. Ол деректерді өңдеу, үлгілерді тану, классификация, регрессия, кластерлеу және басқа да машиналық оқыту алгоритмдерін жүзеге асыру үшін қолданылады.

01

Классификация – шешім ағаштары (Decision Trees), логистикалық регрессия, SVM, нейрондық желілер және т.б.

02

- Регрессия – сызықтық, көпмүшелі (polynomial), Ridge, Lasso регрессиялары
- Кластерлеу – K-Means, DBSCAN, Hierarchical clustering

03

- Өлшемділікті азайту – PCA, LDA
- Модельді бағалау – кросс-валидация, метрикалар (F1-score, accuracy, RMSE)

04

- Машиналық оқытуға арналған деректерді өңдеу – деректерді масштабтау (StandardScaler, MinMaxScaler), ерекшеліктерді таңдау (Feature Selection)

PATTERN

Pattern – бұл Python тілінде жазылған табиғи тілді өңдеу (NLP), веб-скрапинг, машиналық оқыту және деректерді талдауға арналған көпфункционалды кітапхана. Ол әсіресе веб-қосымшалар мен мәтіндік деректерді өңдеу үшін пайдалы.

1. Табиғи тілді өңдеу (NLP)
   - Токенизация – мәтінді сөздер мен сөйлемдерге бөлу
   - Сөз таптарын белгілеу (POS-тегтеу) – зат есім, етістік, сын есім және т.б.
   - Сезімдерді талдау – мәтіннің позитивті, негативті немесе бейтарап екенін анықтау
   - Синтаксистік талдау – сөйлем құрылымын анықтау
   - Лемматизация – сөздердің бастапқы формасын табу
2. Веб-скрапинг
   - HTML және XML талдау (BeautifulSoup сияқты)
   - Google, Bing, Twitter, Wikipedia-дан ақпарат алу
3. Машиналық оқыту
   - Кластерлеу және классификация (k-means, decision tree)
   - Нейрондық желілер мен регрессия әдістері
4. Деректерді визуализациялау
   - Графиктер мен диаграммалар жасау

НАЗАРЫҢЫЗҒА
РАҚМЕТ!