# Порівняння методів векторизації та інструментів NLP

## Основні результати порівняння методів векторизації

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Метод | Переваги | Недоліки | Застосування | Масштабованість |
| Bag of Words | Простота | Не враховує порядок слів | Класифікація тексту | Низька |
| TF-IDF | Знижує вплив частих слів | Може не враховувати контекст | Аналіз тональності | Середня |
| Word2Vec | Зберігає семантичні відношення | Високі вимоги до ресурсів | Нейронні мережі для NLP | Висока |

### Висновки щодо методів векторизації

1. Bag of Words (BoW):  
- Найкраще підходить для простих задач, таких як класифікація тексту або аналіз частоти слів.  
- Ідеальний для невеликих обсягів тексту, де не потрібен аналіз контексту.  
  
2. TF-IDF (Term Frequency-Inverse Document Frequency):  
- Підходить для задач, де важливо враховувати значущість слів у документі відносно всього корпусу, наприклад, аналіз тональності або витягнення ключових слів.  
- Добре працює в середніх за розміром корпусах, проте може втрачати контекст.  
  
3. Word2Vec:  
- Найкраще підходить для задач, що потребують аналізу семантики, таких як рекомендаційні системи, кластеризація текстів або нейронні мережі для NLP.  
- Підтримує високий рівень контекстуального розуміння, але вимагає значних обчислювальних ресурсів.

## Основні результати порівняння інструментів NLP

|  |  |  |  |
| --- | --- | --- | --- |
| Інструмент | Основні можливості | Підтримувані мови | Зручність використання |
| NLTK | Токенізація, стемінг, лемматизація | Багато мов | Середня |
| SpaCy | NER, POS-тегінг | Англійська та інші | Висока |
| Hugging Face Transformers | Підтримка моделей трансформерів | Багато мов | Висока |
| Gensim | Word Embeddings, тематика моделювання | Багато мов | Середня |

### Висновки щодо інструментів NLP

1. NLTK (Natural Language Toolkit):  
- Підходить для основних операцій із текстом, таких як токенізація, стемінг та лемматизація.  
- Ідеальний для новачків у NLP або для академічних проектів.  
  
2. SpaCy:  
- Найкраще підходить для задач, що вимагають швидкості та ефективності, таких як розпізнавання сутностей (NER) та аналіз синтаксичної структури (POS-тегінг).  
- Зручний у використанні та має високу продуктивність.  
  
3. Hugging Face Transformers:  
- Ідеальний для сучасних NLP задач, таких як мовне моделювання, генерація тексту, класифікація тональності.  
- Підтримує широкий спектр передтренованих моделей (BERT, GPT-3).  
  
4. Gensim:  
- Найкраще підходить для задач, пов'язаних із тематичним моделюванням та Word Embeddings.  
- Широко використовується для аналізу великих корпусів текстів.