Практична робота № 8

4 години

Тема: Бібліотека NLTK мови Python**.** Використання нормалізованих даних для аналізу тексту. Векторне представлення.

Мета: Відпрацювати практичні навички написання програм на мові Python для обробки текстової інформації використовуючи методи бібліотеки NLTK.

Література: http://nlpx.net/archives/57

Зміст роботи:

За матеріалами лекції № 5 виконати завдання:

1. Задати текст з 6 речень виконуючи умови: речення повинні мати різну кількість слів; кілька слів мають повторюватись в різних реченнях.
2. Використовуючи заданий текст створити список унікальних слів («мішок слів»
3. Розрахувати частоту кожного терміну в документах.
4. Розрахувати зворотню частотність документів.
5. Отримати статистичну міру оцінки важливості слів в документах.
6. Результати роботи оформити в таблицях у текстовому файлі (.docx). Розрахунки доцільно проводити в MS Excel.
7. Написати програму розрахунку **Term Frequency(TF)** - частоти термінів. Результати вивести на екрен та у csv –file.
8. Написати програму розрахунку **Inverse Document Frequency (**IDF) – зворотня частотність документа, що вимірює важливість терміна в конкретній колекції документів. Результати вивести на екрен та у csv –file.
9. Використовуючи отриману статистичну міру оцінки важливості слова в документі побудувати діаграму за допомогою бібліотеки Matplotlib. Результат вивести на екран та у файл.

Методичні рекомендації.

Модель мішка слів — це спрощене представлення, яке використовується в обробці природної мови та пошуку інформації (IR). У цій моделі текст (наприклад, речення чи документ) представлено як мішок (мультинабір) своїх слів, нехтуючи граматикою та навіть порядком слів, але зберігаючи множинність .

Term Frequency(TF) – частота терміну, яка вимірює наскільки часто зустрічається даний термін в обраному документі. Оскільки в великих документах термін буде зустрічатися більшу кількість раз ніж в маленьких, просто кількість

знаходжень цього слова нам не вистачає. Тому використовують відносну частоту – відношення числа входження слова до загальної кількості слів в документі.

$$TF = \frac{кількість\ появ\ слова\ у\ документі}{загальна\ кількість\ слів\ у\ документі}$$

Inverse Document Frequency (IDF) – зворотня частотність документа, що вимірює важливість терміна в конкретній колекції документів. Деякі слова, наприклад прийменники, зустрічаються в усіх документах дуже часто, хоча майже не мають впливу на сенс тексту. Оскільки під час обрахування частоти терміну, ми вважали кожен токен рівнозначним, нам потрібно зменшити оцінку у словах, які присутні у всіх документах. Для цього і обраховують IDF. Його значення відповідає логарифму від відношення загальної кількості документів в колекції, до кількості документів, в яких присутній обраний термін. Такий розрахунок дозволяє додавати ваги словам, які зустрічаються рідко, на противагу тим, що наявні майже у кожному документі.

$$IDF = log_{10} \frac{кількість\ документів}{кількість\ документів\ з\ заданим\ словом}$$

Отримані значення **TF та IDF** перемножуються для кожного слова і результат використовується у подальшій роботі.
Таким чином ми отримуємо статистичну міру оцінки важливості слова в документі, що є частиною колекції чи корпуса **TF-IDF.**
Приклад:

1. The car is moving on the road.
2. The truck is driving on the highway.

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Контрольні запитання.

1. Поясніть призначення бібліотеки NLTK.
2. Як побудувати список мішка слів?
3. Чи можна для побудови списку мішка слів використати методику токенізації тексту за словами?
4. В чому суть методу Term Frequency(TF)?
5. В чому суть методу Inverse Document Frequency (IDF)?
6. Що є результатом використання частотних методів аналізу текстів?