Изображение выглядит как строительство, зарисовка, черно-белый, силуэт

Автоматически созданное описание

Міністерство освіти і науки України Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського” Факультет інформатики та обчислювальної техніки Кафедра інформаційних систем та технологій

# Лабораторна робота №3

З дисципліни «Технології машинного навчання»

Тема: «Вступ до опрацювання природної мови»

Виконав: Перевірив:

студент групи ІА – 12 пос. Коломоєць С. О. Мельник Михайло Сергійович

Дата здачі

Київ-2024

Мета**:** здобуття студентом навичок реалізації базових методів обробки природної мови, включно з попереднім опрацюванням тексту, формуванням «мішка слів» («bag-of-words»), виділенням стоп-слів і найважливіших слів у документі, створенням тематичних моделей.

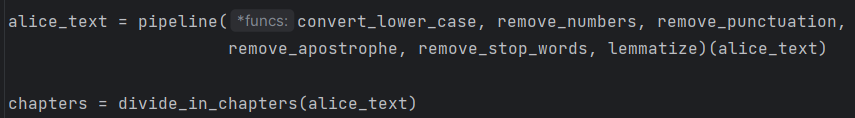
Завдання:

1. Завантажити англійську книгу «Alice's Adventures in Wonderland» <http://www.gutenberg.org/files/11/11-0.txt>
2. Реалізувати пайплайн опрацювання обраного тексту англійською мовою, включно з усією необхідною попередньою обробкою тексту, включно з приведенням слів до нижнього регістру, видаленням стоп-слів, цифр/неалфавітних символів, знаків пунктуації.
3. Розділити текст на глави і в кожній главі відібрати Топ-20 слів за допомогою алгоритму TF-IDF.
4. Реалізувати LDA алгоритм і порівняти результати з отриманими раніше за допомогою TF-IDF. Зробити висновки про застосовність реалізованих підходів.

# Хід роботи

**Виконання завдання:**

1. Було завантажено книгу «Alice's Adventures in Wonderland», текст якої було обоблено кастомним пайплайном, що складається з приведення слів до нижнього регістру, видаленням цифр, знаків пунктуації, апостофів, стоп-слів, і наприкінці застосовується лематизація(перетворення слів до базової форми). Стоп-слова та функція лематизації імпортовані з бібліотеки NLTK(Natural Language ToolKit):

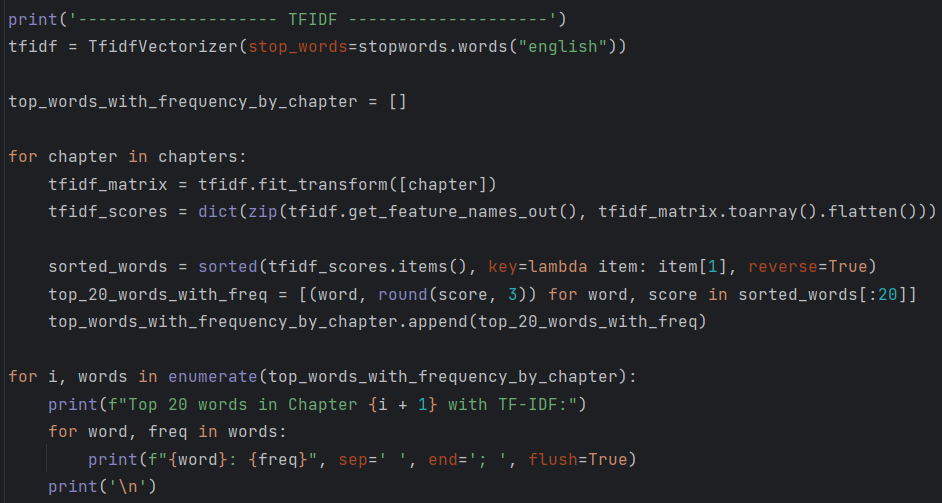


1. Було також розділено текст на глави. Отриманий масив глав береться з 13 індексу тому що так пропускається таблиця «Contents»:





1. За допомогою алгоритму TF-IDF вдбираємо в кожній главі топ-20 слів. Для цього ініціалізуємо TfidfVectorizer, який перетворюватиме кожен розділ на матрицю векторів TF-IDF. Алгоритм TF-IDF дає змогу оцінити «важливість» слова, знизивши вплив низькоінформативних слів і, навпаки, підвивши "вагу" важливих з погляду оцінки приналежності документа до певної тематики.



1. Отримані значення алгоритму TF-IDF:

-------------------- TFIDF --------------------

Top 20 words in Chapter 1 with TF-IDF:

alice: 0.439; little: 0.235; like: 0.173; way: 0.173; see: 0.157; door: 0.141; one: 0.141; think: 0.141; could: 0.125; said: 0.125; thought: 0.125; time: 0.125; eat: 0.11; found: 0.11; get: 0.11; nothing: 0.11; thing: 0.11; well: 0.11; went: 0.11; would: 0.11;

Top 20 words in Chapter 2 with TF-IDF:

alice: 0.397; little: 0.244; mouse: 0.244; said: 0.183; dear: 0.168; go: 0.168; foot: 0.153; thing: 0.153; like: 0.137; must: 0.137; oh: 0.137; thought: 0.137; went: 0.137; cat: 0.122; cried: 0.122; one: 0.122; pool: 0.122; time: 0.122; way: 0.122; could: 0.107;

Top 20 words in Chapter 3 with TF-IDF:

said: 0.534; alice: 0.361; mouse: 0.33; dodo: 0.188; know: 0.173; one: 0.126; soon: 0.11; bird: 0.094; dry: 0.094; long: 0.094; lory: 0.094; prize: 0.094; round: 0.094; thing: 0.094; would: 0.094; could: 0.079; course: 0.079; dear: 0.079; dinah: 0.079; little: 0.079;

Top 20 words in Chapter 4 with TF-IDF:

alice: 0.41; little: 0.304; rabbit: 0.198; one: 0.185; said: 0.185; bill: 0.159; get: 0.119; heard: 0.119; sure: 0.119; thought: 0.119; came: 0.106; made: 0.106; quite: 0.106; thing: 0.106; window: 0.106; go: 0.093; great: 0.093; like: 0.093; moment: 0.093; puppy: 0.093;

Top 20 words in Chapter 5 with TF-IDF:

said: 0.607; alice: 0.409; caterpillar: 0.304; pigeon: 0.14; serpent: 0.14; little: 0.128; well: 0.117; minute: 0.093; know: 0.082; one: 0.082; size: 0.082; think: 0.082; bit: 0.07; egg: 0.07; got: 0.07; last: 0.07; see: 0.07; time: 0.07; tried: 0.07; youth: 0.07;

Top 20 words in Chapter 6 with TF-IDF:

said: 0.494; alice: 0.452; cat: 0.252; like: 0.168; duchess: 0.147; little: 0.147; footman: 0.126; much: 0.126; baby: 0.116; know: 0.116; mad: 0.116; would: 0.116; could: 0.105; thought: 0.105; went: 0.105; get: 0.095; pig: 0.095; see: 0.095; door: 0.084; large: 0.084;

Top 20 words in Chapter 7 with TF-IDF:

said: 0.544; alice: 0.47; hatter: 0.304; dormouse: 0.249; hare: 0.194; march: 0.194; time: 0.157; know: 0.111; thing: 0.111; one: 0.101; well: 0.092; went: 0.092; little: 0.083; say: 0.083; go: 0.065; replied: 0.065; tea: 0.065; asleep: 0.055; could: 0.055; draw: 0.055;

Top 20 words in Chapter 8 with TF-IDF:

said: 0.45; alice: 0.418; queen: 0.396; head: 0.171; king: 0.15; cat: 0.118; three: 0.118; hedgehog: 0.107; like: 0.107; one: 0.107; went: 0.107; began: 0.096; came: 0.096; see: 0.096; soldier: 0.096; two: 0.096; gardener: 0.086; looked: 0.086; away: 0.075; back: 0.075;

Top 20 words in Chapter 9 with TF-IDF:

said: 0.545; alice: 0.497; turtle: 0.258; mock: 0.249; gryphon: 0.191; duchess: 0.182; queen: 0.134; went: 0.124; never: 0.105; little: 0.096; say: 0.096; moral: 0.086; know: 0.077; like: 0.077; quite: 0.077; thing: 0.077; come: 0.067; day: 0.067; make: 0.067; thought: 0.067;

Top 20 words in Chapter 10 with TF-IDF:

said: 0.487; gryphon: 0.335; turtle: 0.335; alice: 0.314; mock: 0.303; would: 0.173; dance: 0.141; lobster: 0.141; soup: 0.13; beautiful: 0.119; join: 0.097; voice: 0.097; could: 0.087; know: 0.087; like: 0.087; whiting: 0.087; come: 0.076; first: 0.076; replied: 0.076; tone: 0.076;

Top 20 words in Chapter 11 with TF-IDF:

said: 0.516; king: 0.353; hatter: 0.285; alice: 0.217; court: 0.204; dormouse: 0.177; one: 0.149; witness: 0.136; queen: 0.122; began: 0.109; rabbit: 0.109; thought: 0.109; white: 0.095; march: 0.081; next: 0.081; voice: 0.081; added: 0.068; cook: 0.068; could: 0.068; first: 0.068;

Top 20 words in Chapter 12 with TF-IDF:

said: 0.647; alice: 0.297; king: 0.285; would: 0.168; little: 0.129; queen: 0.129; jury: 0.116; head: 0.103; know: 0.103; one: 0.103; rabbit: 0.103; white: 0.103; could: 0.091; gave: 0.091; nothing: 0.091; sister: 0.091; must: 0.078; voice: 0.078; dream: 0.065; hand: 0.065;

1. Було реалізовано алгоритм LDA, що розпочинається з токенізація тексту кожного розділу(створення словника LDA). Для моделі LDA було виділено 20 тем, але кількість виділених алгоритмом релевантних тем складає близько 7. Для кожної теми також виводиться 20 найрелевантніших слів з їхніми вагами для наймовірнішої теми.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

1. Отримані значення алгоритму LDA:

-------------------- LDA --------------------

Topic: 17 Chapter: 1

Words: [('said', 0.029), ('alice', 0.027), ('queen', 0.015), ('king', 0.013), ('one', 0.01), ('little', 0.008), ('see', 0.008), ('thought', 0.008), ('like', 0.007), ('head', 0.007), ('hatter', 0.007), ('went', 0.007), ('rabbit', 0.007), ('could', 0.007), ('began', 0.007), ('would', 0.006), ('time', 0.006), ('way', 0.006), ('cat', 0.005), ('looked', 0.005)]

Topic: 4 Chapter: 2

Words: [('alice', 0.017), ('little', 0.017), ('mouse', 0.016), ('dear', 0.014), ('foot', 0.013), ('go', 0.012), ('pool', 0.011), ('oh', 0.011), ('cat', 0.011), ('cried', 0.011), ('must', 0.01), ('thought', 0.01), ('way', 0.01), ('like', 0.01), ('tear', 0.009), ('went', 0.008), ('thing', 0.007), ('swam', 0.007), ('come', 0.007), ('let', 0.006)]

Topic: 6 Chapter: 2

Words: [('alice', 0.017), ('little', 0.017), ('mouse', 0.016), ('dear', 0.014), ('foot', 0.013), ('go', 0.012), ('pool', 0.011), ('oh', 0.011), ('cat', 0.011), ('cried', 0.011), ('must', 0.01), ('thought', 0.01), ('way', 0.01), ('like', 0.01), ('tear', 0.009), ('went', 0.008), ('thing', 0.007), ('swam', 0.007), ('come', 0.007), ('let', 0.006)]

Topic: 13 Chapter: 2

Words: [('alice', 0.017), ('little', 0.017), ('mouse', 0.016), ('dear', 0.014), ('foot', 0.013), ('go', 0.012), ('pool', 0.011), ('oh', 0.011), ('cat', 0.011), ('cried', 0.011), ('must', 0.01), ('thought', 0.01), ('way', 0.01), ('like', 0.01), ('tear', 0.009), ('went', 0.008), ('thing', 0.007), ('swam', 0.007), ('come', 0.007), ('let', 0.006)]

Topic: 6 Chapter: 3

Words: [('said', 0.045), ('alice', 0.039), ('hatter', 0.015), ('know', 0.012), ('dormouse', 0.012), ('mouse', 0.011), ('time', 0.01), ('one', 0.01), ('thing', 0.01), ('march', 0.009), ('hare', 0.009), ('little', 0.008), ('say', 0.007), ('went', 0.007), ('could', 0.006), ('dodo', 0.006), ('well', 0.006), ('long', 0.005), ('go', 0.005), ('like', 0.005)]

Topic: 13 Chapter: 4

Words: [('alice', 0.025), ('little', 0.018), ('rabbit', 0.012), ('said', 0.011), ('one', 0.011), ('bill', 0.009), ('heard', 0.008), ('thought', 0.007), ('sure', 0.007), ('get', 0.007), ('came', 0.007), ('thing', 0.007), ('made', 0.006), ('quite', 0.006), ('window', 0.006), ('voice', 0.006), ('went', 0.006), ('go', 0.006), ('like', 0.006), ('great', 0.006)]

Topic: 15 Chapter: 5

Words: [('said', 0.05), ('alice', 0.03), ('turtle', 0.011), ('would', 0.011), ('gryphon', 0.011), ('mock', 0.01), ('caterpillar', 0.009), ('little', 0.009), ('know', 0.008), ('king', 0.008), ('one', 0.007), ('could', 0.006), ('voice', 0.006), ('tone', 0.005), ('well', 0.005), ('like', 0.005), ('back', 0.005), ('time', 0.005), ('first', 0.005), ('come', 0.005)]

Topic: 9 Chapter: 6

Words: [('said', 0.037), ('alice', 0.034), ('cat', 0.019), ('like', 0.013), ('little', 0.011), ('duchess', 0.011), ('much', 0.009), ('footman', 0.009), ('know', 0.009), ('would', 0.009), ('mad', 0.009), ('baby', 0.009), ('went', 0.008), ('could', 0.008), ('thought', 0.008), ('see', 0.007), ('get', 0.007), ('pig', 0.007), ('thing', 0.006), ('way', 0.006)]

Topic: 6 Chapter: 7

Words: [('said', 0.045), ('alice', 0.039), ('hatter', 0.015), ('know', 0.012), ('dormouse', 0.012), ('mouse', 0.011), ('time', 0.01), ('one', 0.01), ('thing', 0.01), ('march', 0.009), ('hare', 0.009), ('little', 0.008), ('say', 0.007), ('went', 0.007), ('could', 0.006), ('dodo', 0.006), ('well', 0.006), ('long', 0.005), ('go', 0.005), ('like', 0.005)]

Topic: 17 Chapter: 8

Words: [('said', 0.029), ('alice', 0.027), ('queen', 0.015), ('king', 0.013), ('one', 0.01), ('little', 0.008), ('see', 0.008), ('thought', 0.008), ('like', 0.007), ('head', 0.007), ('hatter', 0.007), ('went', 0.007), ('rabbit', 0.007), ('could', 0.007), ('began', 0.007), ('would', 0.006), ('time', 0.006), ('way', 0.006), ('cat', 0.005), ('looked', 0.005)]

Topic: 11 Chapter: 9

Words: [('said', 0.049), ('alice', 0.044), ('turtle', 0.023), ('mock', 0.022), ('gryphon', 0.017), ('duchess', 0.016), ('queen', 0.012), ('went', 0.011), ('never', 0.009), ('little', 0.009), ('say', 0.009), ('moral', 0.008), ('like', 0.007), ('thing', 0.007), ('know', 0.007), ('quite', 0.007), ('thought', 0.006), ('come', 0.006), ('make', 0.006), ('day', 0.006)]

Topic: 13 Chapter: 10

Words: [('alice', 0.025), ('little', 0.018), ('rabbit', 0.012), ('said', 0.011), ('one', 0.011), ('bill', 0.009), ('heard', 0.008), ('thought', 0.007), ('sure', 0.007), ('get', 0.007), ('came', 0.007), ('thing', 0.007), ('made', 0.006), ('quite', 0.006), ('window', 0.006), ('voice', 0.006), ('went', 0.006), ('go', 0.006), ('like', 0.006), ('great', 0.006)]

Topic: 15 Chapter: 10

Words: [('alice', 0.025), ('little', 0.018), ('rabbit', 0.012), ('said', 0.011), ('one', 0.011), ('bill', 0.009), ('heard', 0.008), ('thought', 0.007), ('sure', 0.007), ('get', 0.007), ('came', 0.007), ('thing', 0.007), ('made', 0.006), ('quite', 0.006), ('window', 0.006), ('voice', 0.006), ('went', 0.006), ('go', 0.006), ('like', 0.006), ('great', 0.006)]

Topic: 17 Chapter: 11

Words: [('said', 0.029), ('alice', 0.027), ('queen', 0.015), ('king', 0.013), ('one', 0.01), ('little', 0.008), ('see', 0.008), ('thought', 0.008), ('like', 0.007), ('head', 0.007), ('hatter', 0.007), ('went', 0.007), ('rabbit', 0.007), ('could', 0.007), ('began', 0.007), ('would', 0.006), ('time', 0.006), ('way', 0.006), ('cat', 0.005), ('looked', 0.005)]

Topic: 15 Chapter: 12

Words: [('said', 0.05), ('alice', 0.03), ('turtle', 0.011), ('would', 0.011), ('gryphon', 0.011), ('mock', 0.01), ('caterpillar', 0.009), ('little', 0.009), ('know', 0.008), ('king', 0.008), ('one', 0.007), ('could', 0.006), ('voice', 0.006), ('tone', 0.005), ('well', 0.005), ('like', 0.005), ('back', 0.005), ('time', 0.005), ('first', 0.005), ('come', 0.005)]

1. Порівняння LDA і TF-IDF:

TF-IDF (Term Frequency-Inverse Document Frequency):

* TF-IDF визначає важливість кожного слова в документі на основі частоти його появи в ньому і того, наскільки рідко воно зустрічається в інших документах.
* TF-IDF використовується для знаходження унікальних слів у кожному докумені, що дозволяє оцінити «значущі слова» в контексті документу.
* TF-IDF підходить для завдань класифікації та визначення ключових слів, але не дає інформації про теми в документах.

LDA (Latent Dirichlet Allocation):

* LDA є методом тематичного моделювання, який дозволяє виділити «приховані» теми в текстових даних, де кожна тема – це набір слів з різними вагами.
* LDA представляє кожен документ як суміш тем, і кожна тема характеризується розподілом за релевантними для цієї теми словами.

На відміну від TF-IDF, LDA створює групи тем, які можуть показати взаємозв'язки між словами та документами, що корисно для кластеризації документів за темами або іншими словами – для пошуку і виділення «абстракцій» у документах.

Таким чином, TF-IDF використовується для знаходження важливих слів у кожному розділі, а LDA допомагає виявити теми, до яких відносяться розділи, що дозволяє отримати загальну картину тематики тексту.

Посилання на код: <https://github.com/dEdmishka/MLT/tree/main/lab3>

Висновки: в ході виконання цієї лабораторної роботи я отримав знання щодо базових методів обробки природної мови, включно з попереднім опрацюванням тексту, формуванням «мішка слів» («bag-of-words»), виділенням найважливіших слів у документі, створенням тематичних моделей.