Якименко Д.О.
магістр;
Кулибаба П.О.
магістр,
Черкаський державний технологічний університет

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ МАСИВУ ТЕКСТОВИХ ДОКУМЕНТІВ НА ОСНОВІ ТЕХНОЛОГІЇ TEXT MINING

Бурхливе зростання кількості електронних документів, що спостерігається в даний час, наочно показує, що традиційні механізми обробки електронних документів не спроможні впоратись з потребами користувачів. Таким чином, можна виділити основні проблеми, пов'язані зі збільшенням кількості інформації:

- швидке збільшення обсягу інформації, є причиною труднощів пошуку необхідних документів та організації їх у вигляді сховищ, впорядкованих за змістом;
- більшість технологій роботи з текстовими документами орієнтовані на організацію зручної роботи з інформацією для людини, але практично відсутні можливості для передачі смислового змісту тексту, тобто відсутнє семантичне індексування;
- для ефективного вирішення завдання пошуку необхідно розширити поняття традиційного документа: з документом необхідно пов'язати знання, що дозволяють інтерпретувати й обробляти дані, які зберігаються в цьому документі;
- неструктурована інформація становить значну частину сучасних електронних текстових документів [1, с. 139].

Мета роботи – проаналізувати алгоритм призначений для формування образів документів та розширити його.

Постановка задачі.

Розроблюваний алгоритм формування образів документів заснований на статистичному підході до аналізу текстів на природній мові. Образ кожного документа пропонується формувати у вигляді багатовимірного вектора нормалізованих і зважених одиночних слів (ознак), що зустрічаються в тексті даного документа. Розмірність такого вектора буде дорівнювати кількості унікальних ознак у колекції документів А. Запропонований спосіб формування образів Φ_D складається з таких основних етапів:

$$\Phi_D =< \Phi_P, \Phi_{DP} >, \quad (1)$$

де Φ_P – спосіб вилучення ознак із текстів документів; Φ_{DP} – спосіб відображення документів у просторі їх ознак [1, с. 141].

Вирішення задачі.

Для обрахунку Φ_P необхідно виконати наступні операції:

1. Перетворення початкового документу D до необхідного виду D^* – видалення знаків пунктуації, розмітки документу, однакове для всього документу форматування, застосування алгоритму стремінгу і т.д.

2. Видалення із тексту слів, які не мають сенсу при самостійному використанні – службових слів (прийменники, артиклі, сполучники, частки):

$$\Phi_P^* = D^* - G_{SW}, \tag{2}$$

де G_{SW} – набір службових слів, які заздалегідь визначені.

В результаті, в документі залишаються лише слова, які несуть зміст документу, і утворюють словник документу.

Спосіб відображення документів у простір їх ознак Φ_{DP} заснований на процедурі зважування ознак. Зважування ознак документів пропонується виконувати за допомогою традиційної техніки TF*IDF. Вона використовується для того, щоб з дуже довгих текстів відбиралися тільки слова з максимальною оцінкою важливістю слів у документі, а решта відкидалися. Це дозволяє скоротити обсяг збережених даних. Значимість слова пропорційна кількості вживань цього слова у документі, і обернено пропорційна частоті вживання слова у інших документах колекції.

Першим кроком в обробці текстів є розрахунок ваг TF-IDF для кожного слова ω в кожному документі Φ_P^*:

$$tf(\omega, \Phi_P^*) = \frac{n_{\omega\Phi_P^*}}{n_{\Phi_P^*}}, \tag{3}$$

де $n_{\omega\Phi_P^*}$ – кількість входжень слова ω в документ, $n_{\Phi_P^*}$ – загальна кількість слів в документі;

$$idf(\omega, A) = lg \frac{\left|\Phi_P^*\right|}{\left|(\Phi_P^* \supset \omega)\right|}, \tag{4}$$

де $\left|\Phi_P^*\right|$ – загальна кількість документів у колекції А, а $\left|(\Phi_P^* \supset \omega)\right|$ – кількість документів, у яких зустрічається ω (коли $n_{\omega\Phi_P^*} \neq 0$). Вибір основи логарифму у формулі не має значення, адже зміна основи призведе до зміни ваги кожного слова на постійний множник, тобто вагове співвідношення залишиться незмінним. Таким чином [2],

$$tfidf(\omega, \Phi_P^*, A) = tf(\omega, \Phi_P^*) \times idf(\omega, A). \tag{5}$$

Більшу вагу TF-IDF отримають слова з високою частотою появи в межах документа та низькою частотою вживання в інших документах колекції.

Отже, було запропоновано підхід для оцінювання тематичної близькості документів з використанням зведення ознак і на їх основі зображено алгоритм формування інформаційно-пошукових образів документів, що дозволяє підвищити якість і швидкість виконання автоматичної кластеризації документів. Запропонований алгоритм складається з наступних етапів: створення інформаційних образів документів; зведення ознак документів

документів, тим самим покращується відображення створених інформаційних образів документів.

Список використаних джерел:

1. Оксанич І. Г. Інтелектуальний аналіз масиву текстових документів на основі технології Text Mining / І. Г. Оксанич, Д. М. Піскунов, Д. П. Черниш // Системи обробки інформації. – 2013. – Вип. 2. – С. 139-143.

2. Jones K. S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation : журнал. – MCB University : MCB University Press, 2004. – Т. 60, № 5. – С. 493-502.

Якимів Н.В.

студентка,

Івано-Франківський національний технічний університет нафти і газу

ПЕРСПЕКТИВИ ВИКОРИСТАННЯМ ФРЕЙМВОРКА XAMARIN ПРИ РОЗРОБЦІ МОБІЛЬНИХ ДОДАТКІВ

В наш час, коли інформаційні технології розвиваються з шаленою швидкістю розробка мобільних додатків є необхідною як для великих так і для малих компаній. Найпопулярнішими платформами на даний момент є Android, iOS та Windows Phone, кожна з яких має свої особливості. Тому потрібно врахувати те, що методи розробки під кожну з цих платформ суттєво відрізняються. Тобто, щоб реалізувати мобільний додаток для кожної з них необхідно як мінімум троє розробників, які спеціалізується в певній технології. Даний варіант вимагає багато ресурсів і часу для розробки, а, оскільки, ці платформи швидко розвиваються, то вже готовому додатку потрібна постійна технічна підтримка, що є неефективним для малих компаній. Проте, можна зекономити час та ресурси, написавши кросплатформний мобільний додаток з використанням Xamarin.

Xamarin – це фреймворк, створений для розробки мобільних додатків незалежно від платформи з використанням мови програмування C#. Тобто, використовуючи основні можливості даної мови можна написати мультиплатформний додаток для Android, iOS та Windows Phone не вникаючи в тонкощі розробки певної платформи. Зазвичай спочатку потрібно розробити графічний інтерфейс для Android, iOS та Windows Phone, а потім написати бізнес логіку, яка буде використовуватись на всіх платформах одночасно. Фреймворк Xamarin складається з таких основних частин як:

- Xamarin.IOS – бібліотека класів для C#, що дає розробнику доступ до iOS SDK;
- Xamarin.Android – бібліотека класів для C#, що дає розробнику доступ до Android SDK;