7. Система ипотечного кредитования в Германии / под ред. О. Штекера; [пер. с нем.]. – Дармштадт : Союз немецких ипотечных банков, Ин-т жилища и окружающей среды г. Дармштадта и Земли Гессен, 1997. – 52 с.
8. Спаських Н. М. Зарубіжний досвід розвитку іпотечного кредитування / Н. М. Спаських // Збірник наукових праць ЧДТУ. Серія: Економічні науки. – 2010. – Вип. 21. – С. 72-74.

Бєгун А.В., Білошицький О.В. **УДК 004.91+004.946**

ІЄРАРХІЧНА КЛАСТЕРІЗАЦІЯ ТЕКСТІВ В УМОВАХ ОБМЕЖЕНОСТІ СПОСТЕРЕЖЕНЬ

При розв'язанні різноманітних видів бізнесових і не тільки бізнесових задач часто виникає потреба в динамічному аналізі великої, а іноді величезної кількості текстової інформації для прийняття конкретного рішення в умовах обмеженого часу. Зазвичай така інформація не зберігається централізовано, а, натомість, диверсифікована серед різноманітних джерел (Інтернет, twitter, соціальні мережі тощо).

Розгляд останніх публікацій за цією проблематикою [1-3] дозволяє стверджувати, що задача аналізу текстів, у тому числі семантичного, стає все більш актуальною, так як розв'язувалася у минулому переважно для англомовних текстів. Цей факт вимагає подальшого розвитку існуючих практик розв'язання групи задач, а саме – створення моделей аналізу слов'яномовних текстів, які є більш складними з точки зору граматичного та морфологічного аналізу, що, в свою чергу, накладає велику кількість обмежень на проведення такого лінгвістичного аналізу. З цього приводу подія написання статті з наведеної проблематики безумовно є актуальною.

В якості інструмента дослідження пропонується застосування статистичного пакету R [4] для аналізу блогів в умовах, коли кількість постів в блогах (в подальшому – кількість спостережень) менша за кількість змінних (наприклад, унікальних термінів), які беруть участь в аналізі. Тоді процес моделювання можна представити у вигляді послідовності виконання взаємопов'язаних задач:

- демонстрація найбільш популярних практик аналізу неподібних текстів (кластерізація) в умовах обмеженої кількості спостережень;
- аналіз блогів сайту http://blogs.korrespondent.net, який виступає постачальником вхідної інформації для процесу аналізу з метою ідентифікації найбільш дискусійних тем;
- ідентифікація «слабких» місць при аналізі кирилиці;
- здійснення ієрархічної кластеризації текстів;
- кластеризація методами kmeans та kmedoids і порівняння результатів їх дії.

При виконанні якісного аналізу текстів слід враховувати наступні обмеження:

1) в текстах відсутня первинна категоризація (за часом, тематикою, авторами тощо)[1];

2) до аналізу потрапили 400 останніх постів (за датою додавання), з яких за різноманітними причинами обрано 237 постів; решта є не релевантною для аналізу і виключена з процесу дослідження за такими ознаками: фото-пости, коментарі та пости, що за обсягом складають менш 1Кб.

Із запропонованими обмеженнями процес аналізу можна представити у вигляді декількох етапів.

Eman I. Підготовка даних, що включає в себе добування текстів з блогів, стемінг і очищення текстів, ідентифікація проблем при аналізі російськомовних та україномовних текстів.

Eman II. Аналіз і кластеризація. До основних складових цього етапу слід віднести: нормалізацію (TF-IDF) масиву текстів, ідентифікацію термінів, які часто зустрічаються і корелюються, ієрархічну кластеризацію (hclust), класифікацію за допомогою kmeans та kmedoids.

Eman III. Порівняння.

Якщо дослідити кожний з етапів детальніше, то можна визначити переваги і недоліки окремої складової частини моделі. Так, серед множини методів «добування» текстів було відібрано три основних: HTML parsing, RSS feed, Twitter (@korr.blog). Кожен з цих методів має власні особливості.

Наприклад, для першого методу слід було б написати програмний код для посторінкового кроулінгу HTML з екстрактом текстів, тому що він є досить тривіальним, але не найбільш прозорим; другий метод (аналіз RSS стрічки) є ідеальним в тому випадку, коли існує можливість об'єднання усіх посилань на пости в межах бажаного часового горизонту – для сайту http://blogs.korrespondent.net метод RSS зберігає історію глибиною в одну добу; третій метод для кожного твіта[2] (@korr.blog) при оновленні його в реальному часі має можливість зібрання посилань на пост і експортування тексту за вказівкою (з HTML сторінки посту). Такий підхід забезпечує прозорість та цілісність процесу збору тексту.

Таким чином, сутність третього методу полягає у послідовному виконанні наступних дій:

а) парсинг твітів (**parse tweets**):

- збір твітів;
- добування посилань з тексту твітів;
- видалення неактивних посилань;

[1] Тобто в різні часи окремі теми можуть бути найбільш дискусійними: з грудня 2011 р. до лютого 2012 р. такою темою змогла б бути політична репресія опозиційних сил, з лютого до липня 2012 рр. найбільш популярною темою було обговорення та висвітлення у ЗМІ теми ЄВРО-2012 та їй подібної. Окрім того, окремі автори у своїх блогах висвітлюють найбільш цікаві для них теми (спортивні блоги, соціальні, авторські тощо). В даному дослідженні така первинна категоризація за вказаними параметрами не передбачається.

[2] Твіт – унікальний запис в соціальній мережі Twitter (http://twitter.com).

- добування оригінального посилання;
- валідація посилання;
- видалення дублікатів;

б) очищення текстів (**clean texts**):

- переклад усіх текстів англійською;
- добування оригінального тексту з html коду;
- завантаження тексту до Corpus (векторна змінна в R, елементами якої є тексти);
- перехід до строчного регістру, видалення знаків пунктуації, stop-слів;
- стемінг документів (stemming).

Проблеми при аналізі кириличних текстів (україно- та російськомовних).

В процесі дослідження необхідно перейти від аналізу двомовних текстів (написаних українською та російською мовою) до аналізу одномовних текстів. А тому з ряду причин було вирішено перекласти всі тексти англійською мовою. Основними причинами для такого перекладу стали ряд проблем, а саме:

- відмінювання. Відмінювання слів за відмінками в російській та українській мовах вносить додаткову невизначеність в аналіз, оскільки кожне слово може бути представлене кількома формами. Окрім того, терміни вживаються від різних осіб (від першої особи, другої, третьої), а також в однині і множині. Таким чином, отримуємо додатковий обсяг роботи по коректному стемінгу. А тому прийнято рішення відмовитися від стемінгу російських та українських слів, що значно програє стемінгу англійських термінів;
- проблема малої (рядкової) літери «я» в кодуванні windows-1251 (код 0×FF або 255 у 10-й системі). Вона є «винуватицею» ряду несподіваних проблем в програмах без підтримки чистого 8-го біту [5], що включає пакет R (0×FF означає символ закінчення файлу). Іншими словами, пости можуть бути завантажені в корпус в даному кодуванні лише до першої літери «я» («янукович», «объявлять», «братья»). В останньому слові сенс буде змінений.

Таким чином, кириличні тексти зі спеціальними символами або символами псевдографіки не можуть бути коректно декодовані в кодуванні windows-1251 (вимагають додаткового фільтра unrlencoded, який не підтримується в рішенні R).

Таким чином, необхідно або використовувати інше кодування (наприклад, unicode), що не розв'яже інші супутні проблеми, або ж, з огляду на перераховані вище проблеми, перекласти всі тексти англійською мовою. Для автоматизації перекладу було написано окремий код на R [6]. Даний код успадковує рішення Google Translate, і, використовуючи посилання на джерело посту українською або російською мовою у вигляді, наприклад, http://blogs.korrespondent.net/celebrities/blog/gknbu/a50268 повертає посилання на перекладений англійський текст у вигляді:

http://translate.googleusercontent.com/translate_c?rurl=translate.google.com&sl=auto&tl=en&u=http://blogs.korrespondent.net/celebrities/blog/gknbu/a50268&usg=ALkJrhisevp7b7yg4CxX6_iTDxyBAk4PCQ.

Наведемо приклад фрагменту вихідного тексту російською мовою: «До официального старта Евро остается 150 дней. В разгаре, так называемая, операционная подготовка к Чемпионату. Речь идет о налаживании коммуникаций между принимающими городами, обучении персонала и наведении марафета в целом. Ни для кого не секрет, что, по сравнению с украинцами, поляки получили гораздо больше дивидендов в процессе подготовки к ЧЕ. В первую очередь, речь идет о привлечении немалых ресурсов за счет финансирования из фондов ЕС...».

Перекладений за допомогою R-коду текст матиме наступний вигляд: «Before the official launch of the Euro is 150 days.In the midst of the so-called operational preparation for the championship. It is about establishing communication between the host cities, staff training and marafet hover as a whole. It's no secret that, in comparison with the Ukrainians, the Poles were far more dividends in preparation for the Championship. First of all, we are talking about bringing considerable resources through financing from EU funds…».

Як і очікувалось, переклад не є досконалим, проте це не є критичним в межах даного дослідження. Текст після чистки (переведення в нижній регістр, видалення знаків пунктуації, чисел та видалення зайвих пробілів): «official launch euro days midst called operational preparation championship establishing communication host cities staff training marafet hover secret comparison ukrainians poles dividends preparation championship talking bringing considerable resources financing eu funds...».

Текст після стемінгу[3] буде представлений як «offici launch euro day midst call oper prepar championship establish communic host citi staff train marafet hover secret comparison ukrainian pole dividend prepar championship talk bring consider resourc financ eu fund...». Хоча очевидні деякі проблеми перекладу, вони, звичайно, є незначними при аналізі.

Аналіз тексту та ієрархічна кластеризація.

Після того, як ми підготували текст до аналізу (пости), виконавши його стемінг і завантаживши в корпус, процес аналізу та кластеризації полягатиме в наступному:

- побудувати нормалізовану матрицю (Term Document Matrix) та відфільтрувати рідковживані терміни;
- виконати ієрархічну кластеризацію (побудувати дендограму термінів);
- виконати кластеризацію kmeans та візуалізувати кластери;
- виконати кластеризацію kmedoids та візуалізувати кластери.

[3] Стемінг – процес знаходження основи слова для заданого вихідного слова. Основа слова не обов'язково співпадає із морфологічним коренем слова.

Для нормалізації текстів використовуються наступні основні показники.

TF (term frequency) – частота входжень терміну *t* в документ *d*.

DF (document frequency) – частота входжень документу *d*, що містить термін *t* в корпусі

$$DF(t,D) = \frac{|\{d \in D : t \in d\}|}{|D|},$$

IDF (inverse document frequency)

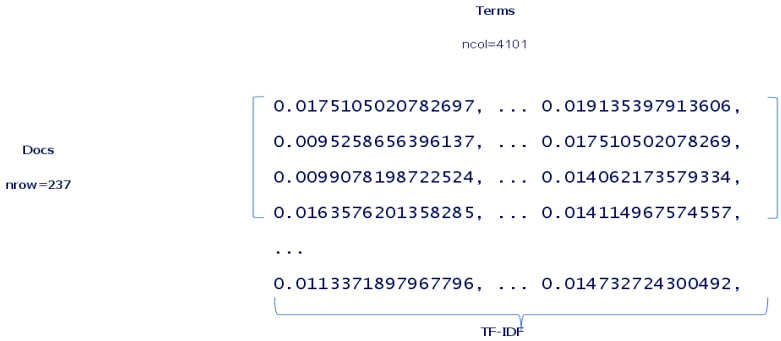
$$IDF(t,d) = \log DF^{-1} = \log \frac{|D|}{|\{d \in D : t \in d\}|},$$

$$TF - IDF(t,d,D) = \frac{TF(t,d)}{DF(t,D)} = \frac{TF(t,d)}{|\{d \in D : t \in d\}|} \times |D| = TF(t,d) \times IDF(t,D)'$$

де $|D|$ – кількість документів в корпусі;

$|\{d \in D : t \in d\}|$ – кількість документів, в яких зустрічається термін *t*.

Таким чином, нормалізація термінів за допомогою TF-IDF надасть більшу вагу термінам з більшою частотою входжень в окремий документ, але низькою частотою в інших документах. В результаті отримаємо нормалізовану матрицю термінів у вигляді:



Таким чином, маємо 237 постів, 4101 унікальний термін (після стемінгу).

Визначення термінів з найбільшою частотою входжень та корельованих термінів

Для розуміння важливості процедури стемінгу приведемо терміни з найбільшою частотою входжень для текстів до проведення стемінгу та після.

Тексти без стемінгу:

> findFreqTerms(dtm, lowfreq=1)

[1] "country" "euro" “european" "government" "internet" "kiev" "kyiv" "money"

[9] "opposition" "party" "people" "political“ "power" "president"

"russia" "social"

[17] "society" "tymoshenko" "ukraine“ "ukrainian" "world" "yanukovych"

Після стемінгу:

> findFreqTerms(dtm, lowfreq=1)

[1] "chang“ "countri" "elect" "euro" "european“ "govern" "internet“ "kiev“

[9] "kyiv" "leader" "money" "opposit" "parti" "peopl" "polit" "power“

[17] "presid" "russia" "russian" "social" "societi" "tymoshenko“ "ukrain" "ukrainian“

[25] "world" "yanukovych"

Як бачимо, після стемінгу результат є більш репрезентативним. Порівняємо тепер терміни, які мають найбільшу кореляцію зі словом «євро» для вихідних текстів та перекладених.

>findAssocs(dtm, 'евр', 0.35) #correlation with term “евро”

евро старт гарант хлеб тыс талисман официальн воплощен будущ чемпионат живет

1.00 0.76 0.74 0.71 0.62 0.55 0.49 0.48 0.35 0.31 0.22

підготовка реплика секрет футбол

0.22 0.22 0.22 0.21

>findAssocs(dtm, ‘euro', 0.35)

euro championship football tourist airport tournament fan poland

1.00 0.68 0.57 0.49 0.45 0.43 0.42 0.42

horribl infrastructur foreign patrol unhappi prepar flashmob

0.38 0.38 0.37 0.37 0.37 0.36 0.35

Як бачимо, англійські терміни мають більші коефіцієнти кореляції та більшу якість (тобто, є більш тематично-орієнтованими на відміну від російських та українських текстів). Відобразимо матриці кореляцій для обох випадків у вигляді графіків (рис. 1 та рис. 2). А тому подальший аналіз буде виконуватись з англійськими термінами.

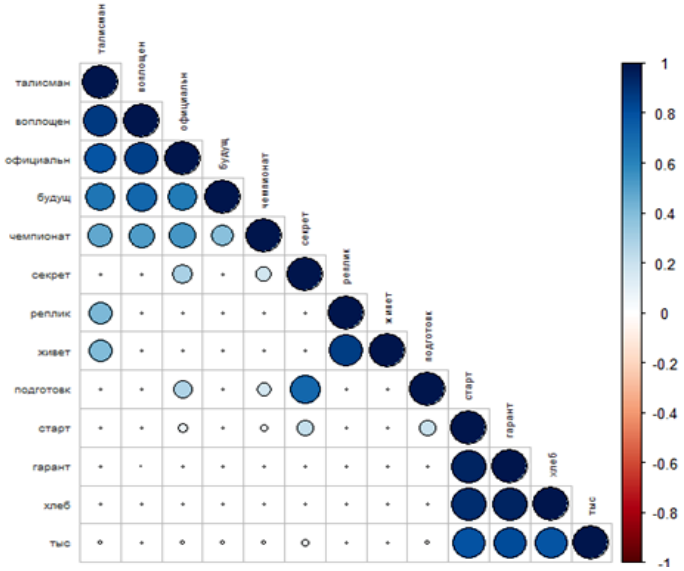


Рис. 1. Матриця кореляції для термінів кирилицею.

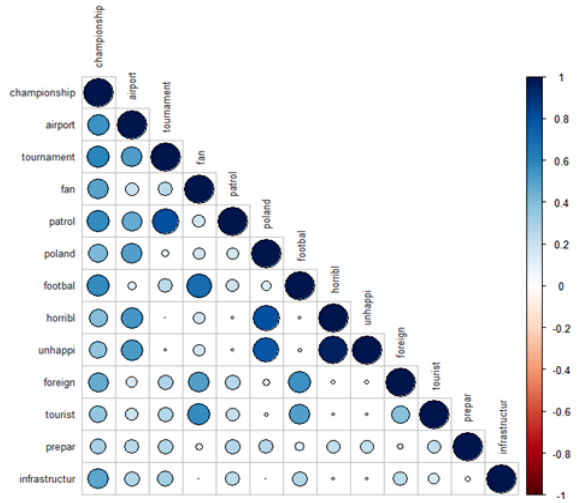


Рис. 2. Матриця кореляції для англійських термінів.

Ієрархічна кластеризація текстів

Ієрархічна кластеризація представляє собою дерево-подібне представлення термінів. При цьому кількість кластерів не задається. Яким чином визначати кластери – залишається на вибір особи, яка виконує кластеризацію. Сутність даного алгоритму полягає в наступному: аналіз виконується на основі множини подібностей для *n* об'єктів (наприклад, термінів, або постів). В даному випадку ми кластеризуємо безпосередньо терміни. Спочатку кожному терміну присвоюється власний кластер, після чого процес виконується ітеративно, на кожному наступному кроці об'єднуються два найбільш подібних кластера, і закінчується, коли усі кластери об'єднані в один. Для знаходження пари найбільш подібних кластерів використовується алгоритм Варда [7], відповідно до якого критерієм для обєднання кластерів на кожній ітерації слугує оптимальне значення цільової функції, що представляє собою середньоквадратичне відхилення (квадрат Евклідової відстані). На кожній ітерації пара кластерів з мінімальною відстанню об'єднується в один кластер.

В результаті проведення ієрархічної кластеризації отримаємо наступний фрагмент (рис. 3).

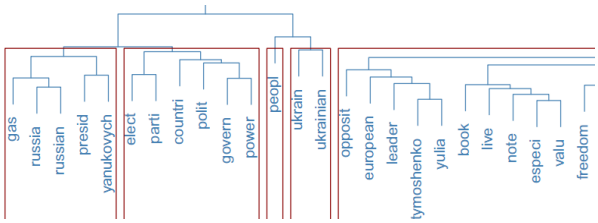


Рис. 3. Результати ієрархічної кластерізації.

Як бачимо, знайдені кластери є досить логічними і кожен з них має свій сенс.

Таким чином, ієрархічна кластеризація є, перш за все, універсальним методом кластеризації за допомогою ряду різних алгоритмів, наприклад, алгоритм цільової функції Варда на основі розрахунку Евклідової відстані між об'єктами. Даний підхід є досить ефективним для великої кількості об'єктів (термінів) в умовах обмеженої кількості спостережень (постів). Окрім того, ієрархічна кластеризація дає можливість зрозуміти, яким чином терміни пов'язані між собою в корпусі, при цьому пакет R забезпечує графічну ілюстрацію у вигляді дерева кластерів.

Виконавши ієрархічну кластеризацію, а також отримавши уявлення про те, яким чином терміни в текстовому масиві корелюють між собою (або не корелюють), ми отримаємо можливість провести більш глибоке подальше дослідження, виконавши кластеризацію методами kmeans та kmedoids з метою розпізнання та групування постів в кластери відповідно до їх змісту.

За результатами подальшої кластеризації можна зробити висновок відносно доцільності проведення окремого дослідження щодо семантичного аналізу текстів, тобто розпізнання, в якому контексті (позитивному, негативному, нейтральному) обговорюється той або інший предмет чи тема.

Джерела та література:

1. Practical Text Miningand Statistical Analysis for Non-structured Text Data Applications / G. Miner, J. Elder, T. Hill, R. Nisbet, D. Delen, A. Fast. – Elsevier Academic Press, 2012.
2. McKnight W. Building Business Intelligence: Text Data Mining in Business Intelligence / W. McKnight // DM Review. – 2005. – P. 21-22.
3. Indurkhya N. Handbook of Natural Language Processing / N. Indurkhya, F. Damerau. – 2nd Edition. – Boca Raton, FL : CRC Press, 2010.
4. [Електронний ресурс]. – Режим доступу : http://www.r-project.org/index.html.
5. [Електронний ресурс]. – Режим доступу : http://uk.wikipedia.org/wiki/Windows-1251.
6. [Електронний ресурс]. – Режим доступу : http://www.slideshare.net/whitish/textmining-with-r.
7. [Електронний ресурс]. – Режим доступу : http://en.wikipedia.org/wiki/Ward's_method.

Боровский Б.И. **УДК 94 (100)**

ПОТРЕБЛЕНИЕ УСЛОВНОГО ТОПЛИВА КАК ПОКАЗАТЕЛЬ ЭФФЕКТИВНОСТИ ЭКОНОМИКИ СТРАНЫ

Введение. Развитие энергетики, наряду с транспортом и связью, ускоряет процесс развития отраслей экономики и всего общества. Энергетические затраты исчисляются в виде условного топлива (ут) или нефтяного эквивалента (нэ). Условное топливо характеризуется теплотой сгорания каменного угля 29,3 МДж/кг, а нефтяной эквивалент - 41,9 МДж/кг.

Анализ литературы. Часто используется экономический показатель - затраты нефтяного эквивалента на единицу создаваемого ВВП [1,2]. Связь между условным топливом и нефтяным эквивалентом следующая: 1 т ут = 0,7 т нэ. Естественно, что увеличение ВВП должно сопровождаться ростом потребления условного топлива. В работе [3] приведены данные по ВВП и потреблению условного топлива для ряда стран и регионов мира в 1990 г.

Основные результаты. В данной статье проведено математическое обобщение литературных данных по ВВП и потреблению ут и получены новые количественные и качественные результаты.

Данные работы [3] приведены в табл. 1

Таблица 1. ВВП и потребление условного топлива.

Страна, регион	ВВП/душу, тыс. долл.	т ут/ душу	ВВП/ т ут, тыс. долл./т	t^0C, среднегодовая температура
Канада	15,1	13,5	1,12	-10,1
США	18,3	11	1,66	2,2
Россия	8,1	8,5	0,95	-10,1
Европа	7,6	4,4	1,72	9
Япония	13,6	5,0	2,72	7
Индия	0,6	0,5	1,2	17,7
Китай	1,1	0,8	1,38	1
Азия (остальная)	1,3	0,52	2,5	11,7
Африка	0,8	0,5	1,6	10,6
Австралия и Новая Зеландия	10,3	7,4	1,39	10,5
Латинская Америка	3,1	1,4	2,21	11,3

В графическом виде данные табл. 1 приведены на рис. 1.