4. Соколов М. Программно-аппаратное обеспечение беспроводных сетей на основе технологии ZigBee/802.15.4. // Электронные компоненты. – 2004. – №12.

5. Умнов А.Л., Головачев Д.А., Ковалев П.А., Шишалов И.С. Система сбора информации с узлов сенсорной сети // Радиотехника. Нелинейный мир. – 2005. – Т.2, № 4. – С. 249–253.

6. Електронний ресурс. – Режим доступу: http://www.maxstream.net/support/downloads.php).

Марина Гвозденко

Володимир Карасюк

(Харків)

ПРОГРАМНІ ЗАСОБИ ЛІНГВИСТИЧНОЇ ЕКСПЕРТИЗИ ЕЛЕКТРОНИХ ДОКУМЕНТІВ

Встановлення авторства тексту (атрибуції тексту) набуває все більшої актуальності у зв'язку зі значним зростанням об'ємів текстової інформації, розповсюджуваних в мережі Інтернет, а також з пов'язаним з цим масовим порушенням авторських прав.

Оскільки підвищення якості та швидкості експертизи по визначенню автора тексту значною мірою залежить від використання засобів інформаційних технологій, то автоматизація лінгвістичної експертизи електронних документів на сьогодні набуває великої актуальності.

Експертні дослідження в галузі атрибуції тексту можуть бути виконані з використанням експертних або формальних методів, а бо з застосуванням обох методів.

Формальні методи визначення автора тексту будуються на виділенні та порівнянні характеристик тексту, що обчислюються.

Саме формальні методи найкраще підлягають алгоритмізації і можуть бути покладені в основу розробки та використання автоматизованих засобів лінгвістичної експертизи.

Існуючі програмні продукти дозволяють враховувати і варіювати різні лінгвостатистичні параметри, різнобічно характеризують текст.

В даний час для атрибуції текстів застосовуються підходи з теорії розпізнавання образів, математичної статистики та теорії ймовірностей, алгоритми нейронних мереж і кластерного аналізу, підрахунок частоти і природи лексичних, орфографічних, синтаксичних і граматичних помилок; дослідження стилістичних факторів письмової мови (довжина слів, довжина речень; кількість складів, приставок і суфіксів на 100 слів); підрахунок відсотка наявних в тексті частин мови: співвідношення дієслів до прикметників, дієслів – до

іменників і т. п., а також показник TTR (TypeTokenRatio) – представлення у формі десяткового дробу співвідношення кількості різних слів із загальною кількістю слів в тексті, метод порівняння гістограм частот слів різної довжини, "наївний байесовський" (НБА) метод, метод розподілу параметра, порівняння кількостей нових слів в суперечливому тексті, метод відносної ентропії, метод стійкості частот, індекс Флеша, FOG-індекс, підхід Колтарда, лінгво-статистичний аналіз неповнозначної лексики, розпізнавання автора тексту з використанням ланцюгів А.А. Маркова, атрибуція Мінімальної Умовної Складності Стиснення (МУСС – атрибуція), апарат машини опорних векторів (SupportVectorMachines, SVM) тощо.

Все більша частина програм визначення автора тексту призначена для експертизи електронних документів і працює в режимі on-line.

На сьогодні, на жаль, відсутні досконалі методики і, відповідно, автоматизовані системи, які дають достовірний та стабільний результат, особливо на навеличких фрагментах тексту, але кожна з нижче розглянутих програм показує досить високий відсоток достовірних результатів лінгвістичних експертиз.

Програма Advego Plagiatus – програма з інтуїтивним інтерфейсом для пошуку в інтернеті часткових або повних копій текстового документа. Advego Plagiatus показує ступінь унікальності тексту, джерела тексту, відсоток збігу тексту. Також програма перевіряє унікальність зазначеного URL. До можливостей Advego Plagiatus належить: перевірка в інтернеті неопублікованих матеріалів, перевірка в інтернеті опублікованого матеріалу, показ сайтів, на яких знайдені збіги, можливість зміни і редагування текстів прямо в самій програмі, при аналізі сторінок сайту доступна функція «прибрати теги». В налаштуваннях програми Адвего Плагіатус можна вибрати пошукові системи, перевірка на унікальність через які найбільше цікавить користувача (Yandex , Google , Yahoo , Nigma , Bing), вибрати розмір шингла, а також задати відсоток збігів, (від 1 до 100) . Аналогічно працює і програма Double Content Finder.

Програма Praide Unique Content Analyser II має широкі можливості з перевірки текстів з використанням пошукових систем. Програма має можливість перевірки скопійованого тексту через буфер обміну і за допомогою імпорту матеріалу з текстового файлу . Також користувач можете перевірити текст вже розташований на веб- сторінці в Інтеренете . У налаштуваннях можна вибрати два способи перевірки – або попассажно (текст розбивається на фрази довжиною від п'яти до десяти слів , які потім шукаються в пошукових системах), або шинглів (матеріал ділиться на фрази довжиною, що дорівнює заданій довжині шингла , «внахлест » , тобто, друге слово в попередній фразі є першим в наступній, і потім також здійснюється пошук в пошукових машинах). Є можливість вибору використовуваних

пошукових систем, програма містить засоби додавання нових пошукових систем. Здійснює перевірку шинглів, довжину яких можна змінювати. Можна задавати кількості слів перекриття шинглів. Виводиться докладний звіт з перевірки в кожній пошуковій системі. У програмі відсутня заміни букв, обробка стоп-слів і немає підтримки роботи з власною базою. Час перевірки за допомогою програми Praide unique content analyzer безпосередньо залежить від введених вами параметрів і розмірів тексту – чим більше тест і менше розмір пасажу або шингли, тим більше час пошуку.Також в налаштуваннях програми є опція захисту IP адреси користувача - між кожним запитом програми до пошукових системах робиться пауза в кілька секунд, щоб при частому зверненні пошукові системи не вирішили , що користувач – робот і не занесли його IP в чорні списки .

Програма Copyscape – застосовується для вияву плагіату, а також підрахунку коефіцієнту унікальності будь-якої сторінки в мережі Інтернет. Copyscape здійснює пошук за індексом Google , або Yahoo , на розсуд сервісу , тобто тут задіяні світові пошукові системи.Сервіс Copyscape має наступні можливості: пошук плагіату на окремих сторінках, що даєрезультати проведення перевірки у вигляді сформованого списку сторінок , на яких буде вказано відсоток збігу , а також візуалізація дублюючого контенту у вигляді колірного виділення на сайті плагіатчіка (від 5 %). У разі появи в мережі Інтернет копій тексту з аналізованого системою сайту , в найкоротші терміни на e-mail адміністратору приходить повідомлення про порушника. Серед наданих можливостей: вказівка доменів дзеркал, з метою подальшого ігнорування інформації, що надходить з них, при проведенні перевірки, можливість вказівки окремого кількості дубльованих пропозицій. Саме за даними параметрами і вважатимуться критерії наявності в тексті плагіату (від 1 до4).

Програма Prostyle (США) здійснює аналіз будь-якого тексту, що вводиться, і виводить в порядку номерів фактори, що дозволяють провести статистичний аналіз значення в будь-яких розбіжностях в двох досліджуваних текстах. Серед факторів, що враховуються програмою Prostyle, знаходяться: граничний індекс чіткості (наскільки даний текст легкий або важкий для розуміння), індекси FOG і Флеша – Кінкейда, показник частотності страждальних конструкцій, що дозволяє достатньо точно виявити індивідуальні особливості автора. кількість використовуваних лексичних одиниць, яка при обчисленні відсотка співвідношення з загальною кількістю слів в тексті дає показник словарного запасу автора, відсоток складних слів по префіксам, суфіксам, кількості складів (у Prostyle – тільки по останньому фактору), середня довжина речення, що прямо корелює з рівнем освіти автора. кількість погрішностей письмового стилю в тексті (можливі помилки: неправильне використання абстрактних іменників; неправильне вживання дієслівних форми прийменників; опущення дієслова; недоречне вживання сленгу і жаргону; використання

застарілих слів; порушення пасивних конструкцій; грубі і непристойні слова; слабке знання мови).

Програма "E'RIDAtextvisor" для аналізу тексту веб-сервером сторінок сайту "E'RIDAtextvisor" проводить загальний аналіз тексту сторінок сайту, аналіз тексту мета-тегів.

Для аналізу тексту сторінки програма сканує текст сторінки. сканує текст в метатегах сторінки, сканує текст анкора (anchor) посилань на сторінці, санує опис посилань.Отримані результати представлені у вигляді: загальна кількість слів на сторінці, а так само окремо в "меті", "тексті" і "посиланнях", загальна кількість символів на сторінці, а так само окремо в "мета", "тексті" і "посиланнях", демонстрація кожного слова з вказівкою кількості однокорінних слів, а так само де і скільки кожне із слів розташовується (у структурних елементах: текст, мета, посилання), відсоткове співвідношення кожного слова до загальної кількості слів, а так само роздільно "% в мета", і "% в тексті" (зручно для контролю ключових слів).

Для аналізу мета-тегів програма сканує текст title, сканує текст description, сканує текст keywords. Отримані результати представлені у вигляді**:** кількість слів в кожному з пунктів, кількість символів в кожному з пунктів, кількість кожного слова, з урахуванням однокорінних слів, відсоткове співвідношення кожного слова в кожному з пунктів (окремо % у title, description і keywords), визначення скільки кожного із слів знаходиться в кожному з пунктів.

Для аналізу посилань програма виконує визначення тексту кожного посилання (anchor), визначення опису кожного посилання.

Отримані результати представлені у вигляді: окремо опис і окремо текст посилання, розбір за словами опису і тексту посилання, визначення кількості слів і кількості символів, як в описі так і в тексті кожного посилання.

Програма eTXT Антиплагіат – виконує перевірку унікальності тексту.

Програма дозволяє провести докладний аналіз унікальності тексту і визначити оригінальність статті у відсотковому співвідношенні. У програмі враховані особливості роботи копірайтера. Програма має 2 версії: установка на комп'ютер користувача і режим on-line.

При роботі зі встановленою програмою користувач може знаходити і виділяти не унікальні фрагменти тексту безпосередньо на відтвореній копії веб-сторінки, що значно полегшує визначення унікальності тексту, створювати докладні звіти перевірки унікальності контенту з можливістю налаштування різних параметрів пошуку – числа вибірок з тексту, кількості слів в шинглі і ін., перевіряти на унікальність всі сторінки сайту, видаючи докладний звіт по сайту, вести пакетну перевірку всіх файлів з теки.

On-line версія програми eTXT Антиплагіат дозволяє: перевірити текст на унікальність незалежно від зовнішніх факторів, таких як швидкість інтернет-з'єднання або встановлена на вашому ПК операційна система, не боятися блокування пошуковими системами, зберігати результати перевірки на сервері і мати можливість надати їх постійну адресу при необхідності.

Програма ШТАМПОМЕР виконує статистичний аналіз тексту і порівняльний аналіз текстів.

При статистичному аналіз тексту програма збирає різні статистичні дані і записує їх у файл результатів у вигляді таблиць відношень або відсоткового змісту: загальні дані, зміст розділових знаків, зміст завершуючих розділових знаків, вміст речень в абзаці, вміст слів в реченні, вміст розділових знаків в реченні. А також таблиць аналізу штампів: повторення штампів n-го рівня, повторення штампів n-го рівня в одному абзаці, повторення штампів n-го рівня в одному реченні.

У цій програмі під штампом n-го рівня розуміємо словосполучення із n слів: штамп 1-го рівня – це одне слово, а 5-го рівня – словсполучення із 5 слів.

При порівняльному аналізі текстів програма обчислює виражені у відсотках різниці відповідних таблиць даних, отриманих на етапі статистичного аналізу текстів. Такі дані характеризують відмінність таблиць, і, чим вище їх значення, тим нижче вірогідність ідентичності авторства початкових текстів.

Іформаційна система "Статистичні методи аналізу літературного тексту" (ІС "СМАЛІТ"). ІС складається з функціонального блоку, призначеного для морфологічного і синтаксичного аналізу текстів, поповнення БД літературних творів, а також внесення виправлень, та аналітичного блоку, що складається з модулів, що реалізують різноманітні методики статистичного аналізу текстів.

Обробка текстів в інформаційній системі проводиться у декілька етапів. На першому кроці виконується автоматизоване розбиття початкового тексту на лексичні одиниці, серед яких виділяються частина (або розділ), абзац, речення, слово. Розбиття здійснюється на основі апарату регулярних виразів. На другому етапі здійснюється автоматична обробка тексту і його морфологічний розбір. На базі побудованого морфологічного розбору проводиться третя стадія обробки тексту – синтаксичний аналіз. На цій стадії для кожнго речення початкового тексту визначаються в середньому близько 15 ознак. Після здійснення обробки вхідного тексту, її результати поміщаються в централізоване сховище (репозиторій текстів, готових для статистичного аналізу).

На наступному етапі користувач може виконувати операції по аналізу текстів, що знаходяться в репозиторії як з використанням клієнтського програмного забезпечення, так і

частково через web, використовуючи інтерфейс, що надається web-вузлом.

Система Плагиат-інформ, розроблена на основі унікальної технології пошуку документів, схожих за змістом, дозволяє легко виявити плагіат, насамперед, в студентських роботах. Масштабування дає можливість об'єднувати декілька вузів в один інформаційний простір, а також факультети, або філії вузу, розташовані на далині один від одного.

Для початку повноцінної роботи з програмою Плагиат-інформ потрібна наявність хоч би одного пошукового індексу.

Створення різних індексів в програмі Плагиат-інформ дає можливість використовувати декілька варіантів пошуку плагіату для рефератів і здійснювати пошук плагіату окремо по кожному індексу.

Спершу пошук плагіату йде по індексу, де реферат, що перевіряється, цілком порівнюється зі всіма документами в індексі. Якщо при перевірці запозичень у файлі не знайдено, то запускається пошук по іншому індексу, в якому документи розбиті на абзаци, і документ, що перевіряється, теж перевіряється по кожному абзацу. Цей пошук виконується повільніше попереднього, зате набагато точніше визначає плагіат і його відсоток. Часто текст, що не був плагіатом після перевірки по першому індексу, стає плагіатом при другому виді пошуку, причому відсоток змісту запозиченої інформації буде достатньо високим. Прграмма дозволяє виконати тестування цілого файлу на наявність запозичень – можна визначити не тільки рівень плагіату, але також можна знайти файл-першоджерело і його зміст. Запозичений з такого документа текст буде відмічений, тестування файлу при перенесенні абзаців усередині тексту, тестування файлу при додаванні нового тексту, видаленні фрагмента, переміщенні речень, тестування файлу, складеного з фрагментів інших документів, тестування файлу, складеного з фрагментів інших документів з перестановкою абзаців.

Програма АТРІБУТОР є лінгвістичним процесором для автоматичного порівняння і класифікації текстів по параметрах індивідуального авторського стилю. Мета роботи програми – розпізнавання автора тексту або видача списку найбільш близьких до нього по стилістиці авторів з числа вхідних в деякий заздалегідь заданий перелік "еталонних" авторів.

При роботі програми передбачено 3 ситуації:

– найбільш вірогідним автором є Х. Цей висновок означає, що в нашій виборці є тексти наданого на дослідження письменника;

– автора цього тексту в нашій базі немає. Цей висновок означає, що присланий текст містить особливості індивідуального стилю, по яких він достатньо різко відрізняється

від наявних у вибірці письменників. Цей текст, мабуть, не містить індивідуальних стилістичних рис;

– список найбільш близьких авторів (в порядку убування вірогідності). Цей висновок означає, що досліджуваний текст по стилістиці не збігається ясно ні з одним з наявних у вибірці письменників і, в той же час, не має різких відмінностей відразу від декількох з них.

Як ознаки для аналізу і оцінки індивідуального авторського стилю використовуються трьохлітерні поєднання – тріади. Обробку проходять всі слова тексту, причому початок і кінець слова доповнюються пропусками, які також враховуються в тріадах. Однакові тріади підсумовуються, із зібраних по тексту тріад виходить профіль, який є пошуковим образом, що характеризує стиль.

Програма ЛІНГВОАНАЛІЗАТОР виконує читання і обробку тексту невідомого походження з метою визначення близькості до одного з авторських еталонів, визначених заздалегідь. "Лінгвоаналізатор" розбирає текст на складові, використовуючи математичну модель, в якій враховані такі характеристики тексту, як число службових слів (прийменників, союзів і інших частинок), морфеми (префіксальні, кореневі, суфіксальні, флективні) і їх послідовності, складність граматичних конструкцій, власне словник, використаний автором. Програма вимірює всі ці параметри і зводить в таблиці, що містять сотні змінних, які характеризують письменника. У кожного автора з бази даних є своя таблиця, яка є авторським еталоном. Початкові тексти "Лінгвоаналізатор" у себе не зберігає.

При введенні аналізованого тексту відбувається побудова ще однієї таблиці по вхідному тексту. Після цього вхідна таблиця зіставляється з **X** таблицями по кожному авторові і виводиться **X** інтегральних величин для оцінки близькості даного тексту до кожного з **X** письменників. Кожна з цих **X** інтегральних величин називається відносною ентропією. Програма повідомить імена трьох авторів, для яких відносна ентропія по даному тексту мінімальна. Таким чином, дослідивши методи і засоби лінгвістичної експертизи, можна зробити декілька висновків:

– проблема визначення авторства тексту є актуальною і увага до неї збільшується у відповідності до збільшення кількості порушень авторських прав у електронному обігу документів;

– методика виявлення авторства не є тривіальною, тому у програмних засобах використовується велика кількість оригінальних евристичних методів;

– отримали розповсюдження методи, основані на фільтрації тексту, стеммінгу, перетворенні символів, що дає змогу системам знаходити запозичені тексти навіть при їх незначній модифікації;

– якість дослідження суттєво зростає при наявності значної бази досліджуваних текстів;

– проаналізовані види статистичного аналізу документа - індекс Флеша, FOG-індекс, стилеметрія, підхід Колтарда, лінгво-статистичний аналіз неповнозначної лексики і топографічний аналіз;

– найточніші дані про автора документа надають статистичні методи авторознавчої експертизи, зокрема стилеметрія.

ДЖЕРЕЛА ТА ЛІТЕРАТУРА:

1. Хмелев Д. В. Распознавание автора текста с использованием цепей А. А. Маркова // Вестник МГУ: Серия 9, Филология. – 2000. – № 2. – С. 115–126.
2. Кукушкина О.В. Определение авторства текста с использованием буквенной и грамматической информации / О. В. Кукушкина, А. А. Поликарпов, Д. В. Хмелев // Проблемы передачи информации. – 2001. – Т. 37. – № 2. – С. 96–109.
3. Баранов А. Введение в прикладную лингвистику: Учебное пособие / А. Н. Баранов. – М.: Эдиториал УРСС, 2003. – 360 с.
4. Галяшина Е. И. Возможности судебных речеведческих экспертиз по делам о защите прав интеллектуальной собственности / Е. И. Галяшина // Интеллектуальная собственность: Авторское право и смежные права. – 2005. – № 9. – С. 50–59.

Ольга Загоруйко
(Біла Церква)

ВИКОРИСТАННЯ ПРОГРАМНОГО КОМПЛЕКСУ NETOP SCHOOL В НАВЧАЛЬНОМУ ПРОЦЕСІ

Озброїти кожну дитину умінням ефективно працювати на комп’ютері, застосовувати у житті набуті знання, уміння та навички – головна мета вчителя інформатики.

Навчання студентів в комп’ютерному класі з використанням традиційних інструментів може виявитися важким завданням. За допомогою інтерактивної дошки і проектора викладач не завжди може ефективно провести практичне заняття. Багатьом знайома картина сьогоднішньої комп'ютерної освіти – вчитель сидить за комп'ютером, група учнів гуртується позаду нього, намагаючись роздивитися, що відбувається на крихітному екрані, або учні працюють за комп’ютерами, а вчитель намагається тримати під контролем кожне робоче місце, перетворюючи себе у контролюючо-обслуговуючий пристрій.