جامعة دمشق

كلية الهندسة المعلوماتية

السنة الرابعة

قسم الذكاء الصنعي

كشف الانتحال باستخدام الويب

الدكتور المشرف:

د. باسل الخطيب

المهندسة المشرفة:

م. أميرة إسبل

تقديم الطلاب:

محمد صالح

مرهف فارس

محمد موسى حمد

فهرس المحتويات

1. ملخص:

بدأ البحث في مجال كشف الانتحال في سبعنيات القرن الماضي، حيث بدأت الدراسات و الأعمال الأولى لكشف التشابه في البرمجة و بالتحديد في جامعات علوم الحاسوب. فعلى مدى الثلاثين عاما الماضية، تم اقتراح عدد كبير من الطرق و الخوارزميات لتحديد التشابه "غير العادي" بين وظائف الطلاب.
و في الآونة الأخيرة، اتجهت جهود الباحثّين والعاملين في مجال اللغات الطبيعية لتحديد أوجه التشابه بين نصوص اللغة الطبيعية، ولكن الأمر لم يكن بالسهل نظراً لغموض وتعقيد اللغات الطبيعية مقارنة باللغات البرمجية.
يوماً بعد يوم، يتزايد الاهتمام بكشف التشابه بين النصوص في العالمين الأكاديمي والتجاري وتتضاعف المواقع و الأنظمة التي تقدم خدمة الكشف عن الانتحال عبر الإنترنت، نبحث في هذا التقرير نظاماً لكشف التشابه بين النصوص المكتوبة باللغات الطبيعية باستخدام محركات بحث على الانترنت.

الفصل الثاني:

مقدمة

2

2. مقدمة:

قسّم المؤرخون و الفلاسفة حياة الإنسان على كوكب الأرض إلى عدة عصور وأطلقوا على كل منها اسماً خاصاً. سميّ عصرنا الحالي *عصر المعلومات* حيث أصبح اكتساب المعارف والحصول على المعلومات أمراً يسيراً جداً، و ذلك بفضل الشبكة العنكبوتية. ولكن مع كل اختراع جديد تأتي سلبيات و مضار جديدة، فتوافر النصوص والمقالات على الشبكة العنكبوتية سهّل كثيراً من عملية الانتحال وسرقة أعمال الآخرين.
أصبحت ظاهرة الانتحال تشكل مشكلة لا يمكن تجاهلها في الوسط العلمي، حيث أظهرت الدراسات مؤخراً أن 40% من الطلاب اعترفوا بأنهم قد قاموا بعملية نسخ حرفي مرة واحدة على الأقل، وأظهرت الدراسة ذاتها أن 70% من الطلاب لم يعتبروا ذلك غشاً [10].

2.1 تعريف الانتحال - Plagiarism:

في اللغة العربيّة:

- "النِّحْلةُ: الدَّعْوَى. وانْتَحَل فلانٌ شِعْر فلانٍ. وتَنَحَّلَه: ادَّعاه وهو لغيره." [لسان العرب]
- "وانْتَحَلَهُ وتَنَحَّلَهُ: ادَّعَاهُ لنَفْسِهِ وهو لغيرِهِ." [القاموس المحيط]

و عرّف كل من Mike Joy و Michael Luck الانتحال عام 1999 [3]على أنّه:

“Unacknowledged copying of documents or programs” that can “occur in many contexts: in industry a company may seek competitive advantage; in academia academics may seek to publish their research in advance of their colleagues.”

كما عرّف Stuart Hannabuss (2001) الانتحال[3]:

“Unauthorized use or close imitation of the ideas and language/ expression of someone else and involves representing their work as your own.”

ونعرف الانتحال باختصار بأنه:

إعادة استخدام شخص لكتابات و أفكار أشخاص آخرين –جهد الآخرين بشكل عام- و نسبها لنفسه سواء بشكل مباشر أو غير مباشر (شكل غير مباشر عدم ذكر المصدر أو اسم الكاتب مثلاً.)

2.2 أشكال الانتحال:

يمكن للانتحال أن يكون بأشكال متعددة منها (Martin,1994)[3]

1.1.1 *Word-for-word plagiarism*: النسخ الحرفي لنص ما، مقاطع، جمل منشورة مسبقاً دون الإشارة إلى المؤلف الأصلي.

1.1.2 *Paraphrasing plagiarism*: تغيير بالكلمات أوالمفردات بحيث يبقى المضمون نفسه و يبقى قابل للتميز.

1.1.3 *Plagiarism of the form of a source*: نسخ هيكلية مناقشات وبراهين في نص آخر مع بعض التغيير بالمفردات.

1.1.4 *Plagiarism of ideas*: عملية إعادة استخدام الأفكار من دون استخدام مفردات المصدر.

1.1.5 *Plagiarism of authorship*: عملية أخذ عمل شخص آخر كاملاً و وضع الاسم عليه كمؤلف له.

كما تجدر الإشارة إلى نوع الانتحال الذي يسمى Ghost-writing حيث يقوم الشخص فيه بتوظيف شخص آخر لكتابة مقالات، وظائف، إلخ... تنشر باسمه، و هذا النوع يصعب جداً كشفه. و قد انتشرت مؤخراً مواقع على الانترنت تقدم خدمة كتابة المقالات للطلاب و سميت هذه المواقع بـ *Paper mills* مثل موقع www.essaymill.com.

2.3 أمثلة عن حالات انتحال:

النص الأصلي:	***"Globalization means that events in one part of the world have ripple effects elsewhere, as ideas and knowledge, goods and services, and capital and people move more easily across borders. Epidemics never respected borders, but with greater global travel diseases spread more quickly. Greenhouse gases produced in the advanced industrial countries lead to global warming everywhere in the world. Terrorism, too, has become global. As the countries of the world become more closely integrated, they become more interdependent. Greater interdependence gives rise to a greater need for collective action to solve common problems*** ***Joseph E. Stiglitz (2006), Making Globalization Work. London: Penguin Books, p. 280."*** **[8]**
Globalization means that events in one part of the world have ripple effects elsewhere, as ideas and knowledge, goods and services, and capital and people move more easily across borders. As the countries of the world become more closely integrated, they become more interdependent.[8]	
"Globalization means that events in one part of the world have ripple effects elsewhere, as ideas and knowledge, goods and services, and capital and people move more easily across borders. As the countries of the world become more closely integrated, they become more interdependent. (Stiglitz 2006, p. 280)."[8]	
"You might say that epidemics never respected borders. But nowadays, with greater global travel, these diseases spread more quickly. Greenhouse gases produced in a certain country lead to global warming everywhere in the world. Terrorism, too, has become global. Therefore, we can say globalization means that events in one part of the world have ripple effects elsewhere, as a result of ideas and knowledge, goods and services, and capital and people moving more easily across borders."[8]	

الفصل الثالث:

الأعمال ذات الصلة

3

3. الأعمال ذات الصلة:

Turnitin	
الموقع	**www.turnitin.com**
السنة	1996
خوارزمية المقارنة	[3] Approximate string matching
نوع النظام	Web-based
اللغات المدعومة	Natural languages
التكلفة	في السنة 3,000$
ملاحظات	أشهر نظام كشف انتحال انتاج شركة iParadigms المحدودة المسؤولية. يقارن النظام مع فهرس خاص به لمحتويات الإنترنت و قاعدة معطيات ضخمة تتضمن أكثر من 125 مليون مقالة.[1]
JPLAG	
الموقع	**www.ipd.uni-karlsruhe.de/jplag**
السنة	1997
خوارزمية المقارنة	Overlap of longest common substrings [3]
نوع النظام	Web-based [11]
اللغات المدعومة	Java, C#, C++, Scheme, and natural languages
التكلفة	مجاني
MOSS	
الموقع	**theory.stanford.edu/~aiken/moss**
السنة	1994
خوارزمية المقارنة	Winnowing algorithm[11]
نوع النظام	Web-based
اللغات المدعومة	C, C++, Java, Javascript, Pascal, Ada, Lisp, Python, C#, Perl
التكلفة	مجاني
Sherlock	
الموقع	**www.cs.su.oz.au/~scilect/sherlock**
السنة	1994
خوارزمية المقارنة	Incremental comparison of two files
نوع النظام	Java application
اللغات المدعومة	Programming languages and natural languages
التكلفة	مجاني و مفتوح المصدر
SNITCH	
الموقع	**----**
السنة	2005
خوارزمية المقارنة	[3] Approximate string matching
نوع النظام	Java application
اللغات المدعومة	Natural languages
التكلفة	في السنة 3,000$

أنظمة أخرى

مجانيّة: **CopyTracker**، **Plagium**، **SeeSources**، **eTBLAST**، **Chimpsky**

تجاريّة: **Plagiarismdetect**، **Copyscape**، **Ephorus**، **Plagiarism-detector**

الفصل الرابع:

البنية العامة للنظام

4

4. البنية العامة للنظام

4.1 توصيف النظام:

بشكل عام العوامل التي تحدد نظام كشف الانتحال بدقة هي [9]:

1) مجال البحث 2) زمن التحليل 3) عمق الاختبار 4) خوارزميات المقارنة 5) الدقة

1) Scope of search 2) Analysis time 3) Check intensity 4) Comparison algorithm type 5) Precision

فيما يلي شرح لهذه العوامل بشكل عام و تحديدها بالنسبة لنظامنا (خصائص النظام بالخط العريض):

مجال البحث **Scope of search**	مجال البحث يمكن أن يكون: الانترنت، باستخدام محركات البحث، قواعد معطيات خاصة محلية، أو على مستوى المؤسسات. يقابل هذا التقسيم النوعيين التاليين: *Open-system, hermitic-system* **Internet (Open-system) باستخدام محركات البحث***
زمن التحليل **Analysis time**	الزمن بين لحظة تقديم المستند للفحص و لحظة ظهور النتائج. **تم حساب زمن التحليلي بالتفصيل في الفصل الثامن، الاختبارات.**
عمق الاختبار **Check intensity**	كيفية تقسيم المستند (فقرات، جمل، كلمات..) و تردد البحث الذي يقوم به النظام عن أقسام المستند. **عمق الاختبار متغير في النظام، حيث يحدد المستخدم العمق الذي يريده. يتحدد العمق بعدد الجمل في المقطع الواحد** **عدد الجمل في المقطع = عدد جمل النص. يعامل النص كله كمقطع، عمق الاختبار أصغري** **عدد الجمل في المقطع = 1 .تعامل كل جملة على أنها مقطع بحد ذاتها، عمق البحث أعظمي**
خوارزميات المقارنة **Comparison algorithm type**	خوارزمية المقارنة المستخدمة، خوارزميات إحصائية، خوارزميات تعتمد على المعنى.... **Fingerprint-based system حيث طبقنّا خوارزميتي Winnowing and LC ****
الدقة **Precision**	دقة النظام، نظام ذو دقة عالية أي نظام يكون عدد النتائج الإيجابية الكاذبة و السلبية الكاذبة قليل جداً، كما يمكن نقارن بعدد الكلمات أيضاً في المستند. **تم حساب دقة النظام في الفصل الثامن، الاختبارات.**

*لما نستخدم الـ Open-system؟

جميع برامج كشف تشابه النصوص التي لا تستخدم الانترنت كفضاء بحث تكون محدودة، فإن بعض محركات البحث، Google مثلاً، لا تفهرس صفحات انترنت فقط و إنما ملفات PDF و Word و مواقع أرشفة مقالات كموقع Citeseer، هذا كان من جهة، و من جهة أخرى صعوبة بناء قاعدة معطيات ضخمة تتضمن عدداً كبيراً من المقالات دفعنا لاستخدام محركات البحث. أما سلبيات و مشاكل محركات البحث فسنأتي على ذكرها لاحقاً.

**لماذا Fingerprint-based system؟

يبين الشكل التالي تعقيد الخوارزميات الاحصائية و الخوارزميات التي تعتمد على المعنى. نلاحظ من الشكل أن الخوارزميات الإحصائية (Fingerprinting) تعقيدها أقل بكثير من تعقيد الخوارزميات التي تعتمد على المعنى (Tree matching) لكنها بالمقابل تعطي نتائج أقل دقة. في النظام تم التركيز على سرعة التنفيذ و تُرك للمستخدم مهمة التأكد من نتائج النظام فلا يمكن الاعتماد على الآلة اعتماداً كلياً في كل زمان و مكان.

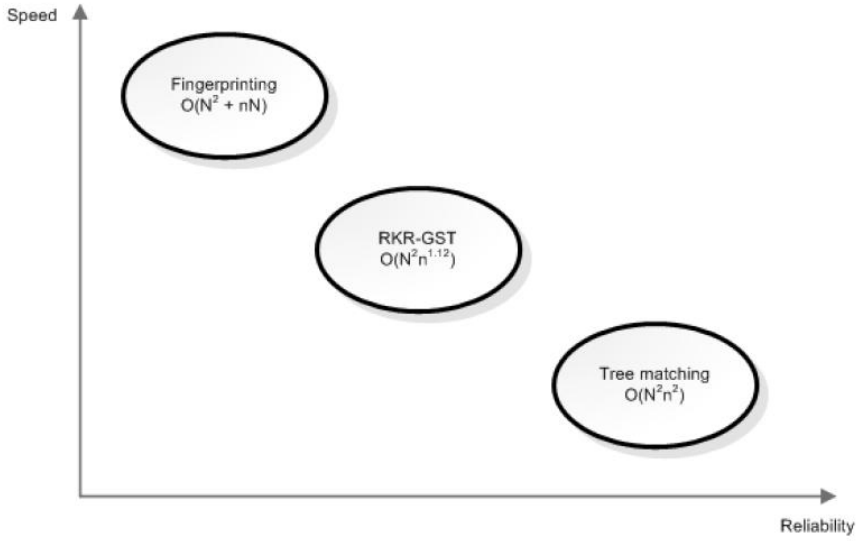
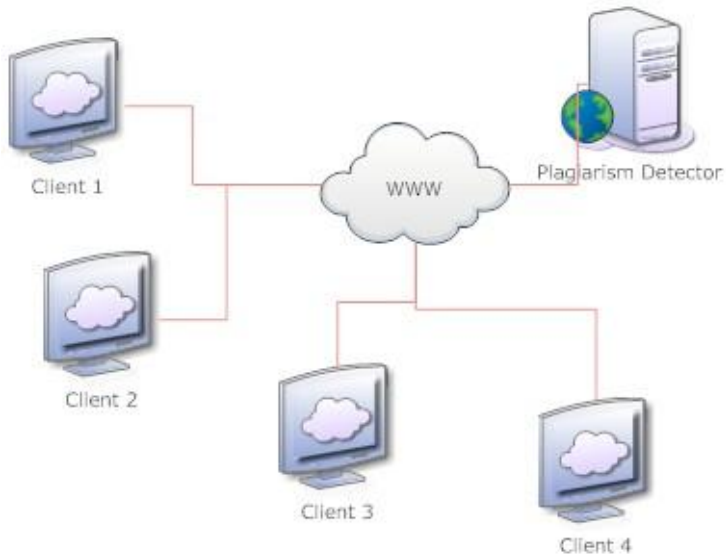
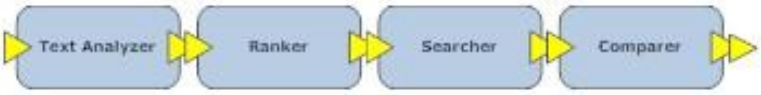


Fig. 6.1. Speed and reliability of different plagiarism detection schemes [4]

النظام عبارة عن خدمة تقدم عن طريق الانترنت، حيث يدخل الزبائن إلى موقع النظام و يقدمون مستندات للمعالجة توضع في رتل المعالجة، ثم يقوم النظام الموجود على المخدم بتفحص المستندات واحداً تلو الآخر.



4.2 الوحدات الرئيسية في النظام:



يتكون النظام من الأجزاء **الرئيسية** التالية:

2. **محلل النصوص Text Analyzer**
3. **مرتّب الجمل Ranker**
4. **الباحث Searcher**
5. **المقارن Comparer**

في مايلي شرح مفصّل لكل من هذه الأجزاء:

4.2.1 محلّل النصوص Text Analyzer:

يتكون محلّل النصوص من جزئين رئيسيين هما: **1-)** قارئ الملفات **2-)** المحلّل اللغويّ Parser

1-) قارئ الملفات: عبارة عن وحدة برمجية لكي نستطيع الحصول على محتوى ملفات من أنماط عديدة. سنذكر الأدوات المستخدمة لقراءة الملفات في فقرة أجزاء خارجية. و الآن يكفي أن نقول أن النظام يعالج ملفات من أنماط: PDF, Word, HTML, and Txt

2-) المحلّل اللغويّ Parser: دخله ملف نصي غير مهيكل (Plain text) و خرجه سلسلة من الجمل و كل جملة تكون عبارة عن سلسلة من الكلمات.
هذا الجزء من النظام تمّ باستخدام أداة مجانية و مفتوحة المصدر هي احدى أدوات مشروع OpenNLP الذي يعتبر مظلة لمعظم المشاريع المفتوحة المصدر لمعالجة اللغات الطبيعية [17]. يوجد حالياً العديد من الأدوات و المشاريع تحت اسم OpenNLP استخدمنا منها OpenNLP Tokenize و تم تعديلها بما يتناسب مع متطلبات النظام.

4.2.2 مرتّبّ الجمل Ranker:

يتميز النظام بأنه يقوم بالاختيار التلقائي للجملة التي يجب البحث عنها على الانترنت، بينما معظم الأنظمة التي تستخدم الانترنت كمجال بحث تطلب من المستخدم ادخال جملة البحث. و لنقوم بتنفيذ الاختيار التلقائي للجملة الأفضل كان من الواجب ترتيب الجمل حسب أهميتها، نقصد بأهمية الجملة كمية المعلومات التي تحملها هذه الجملة عن النص ككل. هناك عدة طرق للترتيب منها:

1. تحديد أهيمة الجملة حسب عدد الكلمات فيها[2].
2. تحديد أهيمة الجملة حسب عدد الكلمات المفيدة أي باستثناء الـ Stop words مثل أحرف الجر، أدوات التعريف و الضمائر[5]
3. تحديد أهيمة كلمات النص و من ثم تحديد أهيمة الجملة حسب أهيمة كلماتها.

تختلف استراتيجيات تحديد أهمية الكلمة في النص منها:

1) تحديد أهيمة الكلمة حسب عدد مرات ورودها في النص فالكلمة الأقل وروداً تحمل قيمة أكثر[6]
2) تحديد أهمية الكلمة بتابع يعتمد على تابع التوزع الطبيعي

الآن لنناقش كل من التوابع السابقة على الفقرة النصية التالية:

"Basically the Winnowing algorithm works in the same way as described in section 2.2.1, but separates itself from the rest by the way it selects the fingerprints. Other algorithms select a number of the hashed n-grams as fingerprints for the documents by using the 0 mod p approach. In this way they may leave large gaps between the fingerprints and thereby allow copied parts to go undetected. To avoid this, Winnowing works with a window of consecutive n-grams. By making their algorithm select at least one fingerprint from each window, Winnowing can guarantee to detect at least one n-gram in a shared substring of length w + n - 1 or longer ([ASW03] section 1).

The Winnowing algorithm will always select the minimum hash value from each window. This is done to increase the chance of matches being caught. The Winnowing algorithm works with a guarantee threshold t. If there exist a shared substring longer than t, a match is guaranteed. The algorithm also works with a noise threshold k, meaning that winnowing will not detect matches shorter than k. A small value of k will therefore, potentially, find a lot of false positives2. A larger value will find less false positive but might not detect some true positives. Moreover, a large value of k makes it harder to detect reordering of code, since substrings smaller then k will not be detected. A theoretical and experimental analysis of this trade-off is also given in [ASW03]"

1. أهمية الجملة حسب عدد كلماتها *Words Count*:

$$v(s) = a.x$$
$$where\ s: sentence,\ x: number\ of\ word\ in\ s$$
$$a: constant\ \&\ a > 0$$

تزداد أهمية الجملة بازدياد عدد كلماتها. التابع سهل التنفيذ لكنه غير دقيق فإذا طبقناه على المقطع النصي السابق تكون الجملة التالية هي الأهم:

"By making their algorithm select at least one fingerprint from each window, Winnowing can guarantee to detect at least one n-gram in a shared substring of length w + n - 1 or longer ([ASW03] section 1) . "

البحث عن هذه الجملة باستخدام Google لا يعطي أي نتيجة.

2. أهمية الجملة حسب عدد كلماتها المفيدة *Words Count Without Stop-Words*:

$$v(s) = a.x$$
$$where\ s: sentence,\ x: number\ of\ word\ in\ s\ \&\ word\ \notin \text{Stop Words}$$
$$a: constant\ \&\ a > 0$$

تزداد أهمية الجملة بازدياد عدد كلماتها المفيدة. أدّق من التابع السابق لكنه لا يأخذ بعين الاعتبار تكرار الكلمات المشكلة مما يجعله يعطي نتائج غير مرتبطة أحياناً.

و بتطبيق التابع على الفقرة النصية تكون الجملة الأهم:

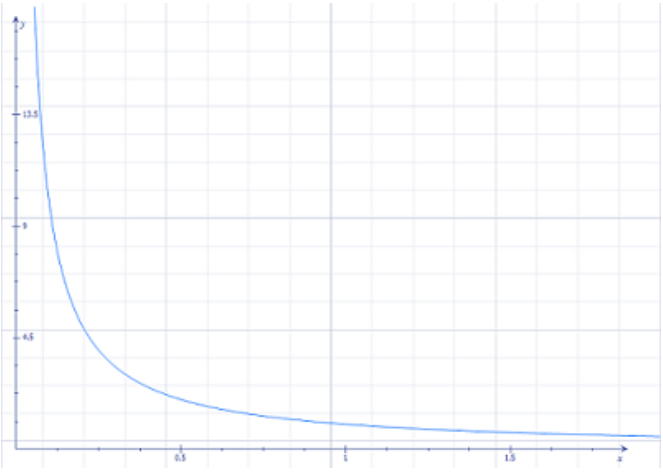
"Basically Winnowing algorithm works way described section 2.2.1 separates rest way selects fingerprints "

يعطي البحث نتيجة واحدة مرتبطة بالموضوع بشكل مباشر من عشر نتائج.

3. أهمية الجملة حسب أهمية كلماتها *Word Importance*:

$$v(s) = Avg(value(w))$$
$$where\ s: sentence,\ w: word\ \in s$$

التابع يعتمد على value(w) : الاستراتيجية المتبعة لتحديد أهيمة الكلمة في النص



i. **تحديد أهيمة الكلمة حسب عدد مرات ورودها في النص فالكلمة الأقل وروداً تحمل قيمة أعلى**

التابع يعطي للكلمة الأقل وروداً القيمة الأعلى. بعد الاختبار وجدنا أن التابع جيد لكن ليس كالمطلوب فمثلاً يمكن أن يرد في النص اسم شخص مرة واحدة فيسند التابع لهذا الاسم القيمة الأعظم!

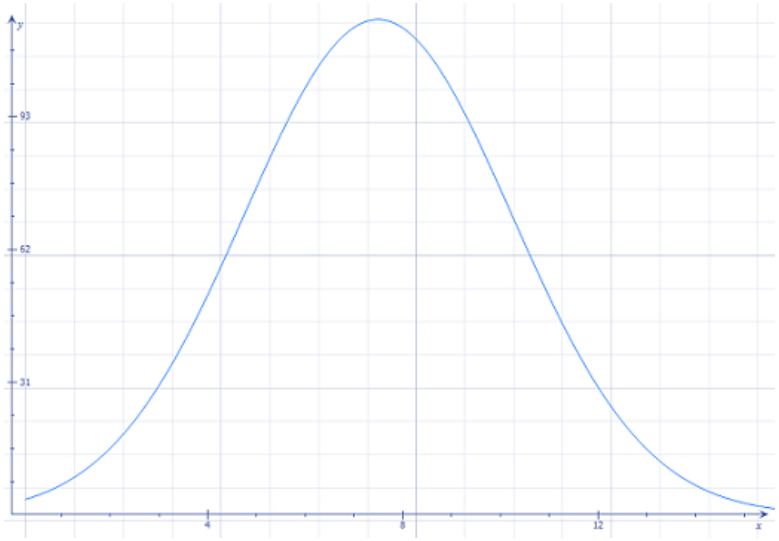
لنوضّح أكثر، بفرض لدينا مقالة عن الاحتباس الحراري و التغيّر المناخي و ورد في المقالة اقتباس للأمين العام للأمم المتحدة

ثم اسمه، سيعطي التابع في هذه الحالة قيمة عالية لكلمات "بان كي مون" ولكن ما العلاقة بين "بان كي مون" و "الاحتباس الحراري" أو "التغير المناخي"!
ولكن بالمقابل يعطي التابع نتائج جيدة على بعض النصوص، كالمثال الذي اختبرنا عليه التابعين السابقين، نتيجة تطبيق التابع هي:

"The algorithm also works with a noise threshold k, meaning that winnowing will not detect matches shorter than k "

وردت كلمة Noise مرة واحدة في النص و بالتالي اختار التابع الجملة السابقة كأهمّ جملة، وبالبحث باستخدام Google كانت النتيجة الأولى هي المقالة المطلوبة لحسن الحظ!

ii. تحديد أهمية الكلمة بتابع يعتمد على تابع التوزع الطبيعي



التابع ممثل بالشكل المجاور و ذلك اعتماداً على تابع التوزع الطبيعي.
في هذا التابع الكلمات التي تكرر كثيراً أو ترد بشكل نادر تكون قليلة الأهمية أما الكلمات التي تكرر بشكل متوسط تكون مهمة.

سنشرح الآن كيف مثلنا التابع و معادلاته التي تتعلق بكل نص:

أولاً، معادلة تابع التوزع الطبيعي هي كالتالي:

$$f(x) = e^{a.x^2+b.x+c} \quad (*)$$

الثوابت

- a يحدد عرض قمة التابع أو ما يسمى بالجرص
- b يحدد انزياح الجرص (القمة) على محور الـ x والذي يمثل عدد التكرارات
- c يمثل ارتفاع الجرص

نحن نريد أن تكون قمية التابع عند أقل تكرار (minRepetition) و أكثر تكرار (maxRepetition) أقل ما يمكن، و عند ((minRepetition+maxRepetition)/2) أعلى ما يمكن و بالتالي:

$$f(\text{minRepetition}) \to 0$$
$$f(\text{maxRepetition}) \to 0$$
$$f\left(\frac{\text{maxRepetition} + \text{minRepetition}}{2}\right) = c$$

نعوض في المعادلة (*) ونحل المعادلات نحصل على العلاقات التالية:

$$a = \frac{\omega - n.b - c}{n^2}$$
$$b = c.\left(\frac{m^2 - n^2}{n^2.m - n.m^2}\right) + \omega.\left(\frac{n^2 - m^2}{n^2.m - n.m^2}\right)$$

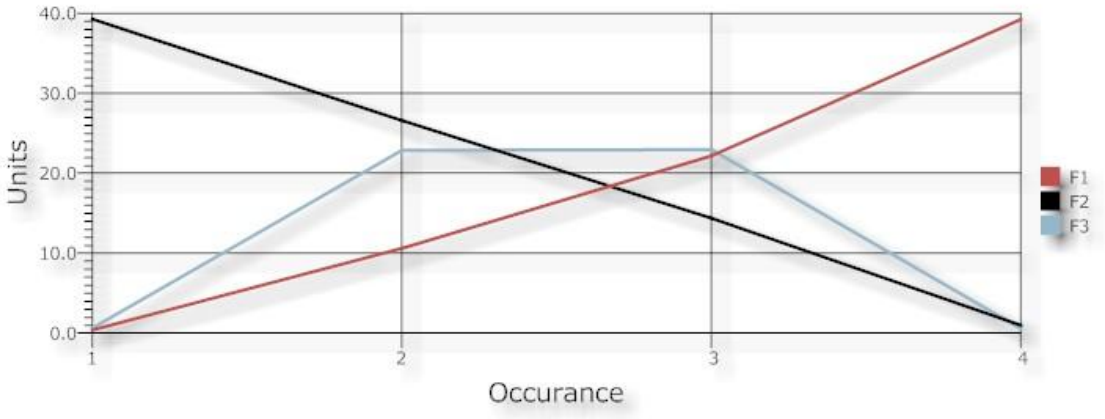
$$c = \omega.\left(\frac{z.u.n^2 - z^2.u.n - z^2}{n^2 + u.z.n^2 - z^2 - z^2.u.n}\right)$$

$$\omega = -6.907755 \approx \ln(0.001)\,, n: minRepetition, m: maxRepetition,$$

$$u = \frac{m^2 - n^2}{n^2.m - n.m^2}$$

$$z = \frac{m+n}{2}$$

نحدد قيم الثوابت باستخدام المعادلات السابقة بالنسبة لكل نص.



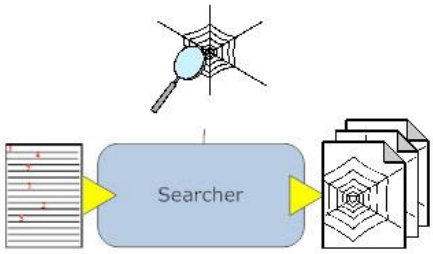
مقارنة توابع ترتيب الجمل

F1: أهمية الجملة حسب عدد كلماتها المفيدة

F2: تحديد أهيمة الكلمة حسب عدد مرات ورودها في النص فالكلمة الأقل وروداً تحمل قيمة أعلى.

F3: تحديد أهمية الكلمة بتابع يعتمد على تابع التوزع الطبيعي

4.2.3 الباحث Searcher:



يتولى هذا الجزء من النظام مهمة البحث على الانترنت. يتم البحث عن الجمل المرتبة باستخدام واجهات التطبيق البرمجية لمحركي البحث Google و Yahoo. فيما يلي شرح بسيط لكل من الواجهتين

4.2.3.1 واجهة بحث Yahoo البرمجية (Yahoo Search API) :

توفر شركة Yahoo العديد من واجهات التطبيق البرمجية (APIs) للعديد من التطبيقات ومنها واجهة البحث (Yahoo Search API) التي تتيح للمطورين استخدام ميزات البحث في Yahoo ضمن مواقعها وتطبيقاتهم عن طريق استعلام مخدم Yahoo الخاص بالبحث ثم معالجة نتائج البحث.

جميع خدمات Yahoo تستخدم آلية REST (Representational State Transfer) والتي هي آلية للتخاطب بين الزبون و المخدم وبالتالي تستخدم الطلب HTTP GET من أجل طلب المخدم وتمرر له مفتاح التطبيق (Application key) والذي يتم طلبه من شركة Yahoo بالإضافة إلى معاملات أخرى تخص البحث المراد.

طلب البحث مبني حسب القواعد المصرّح بها من قبل Yahoo ونتائج البحث أيضا مبنية حسب قواعد Yahoo.

مثال الطلب التالي: `http://api.search.yahoo.com/WebSearchService/V1/webSearch?`

تبدأ باسم المضيف ثم اسم الخدمة المطلوبة (WebSearchService) بعدها رقم النسخة ومن ثم (`V1`) الإجراء الذي سوف يتم استخدامه(`webSearch`).

ومن هنا نستطيع أن نضيف خيارات البحث اعتمادا على قواعد ياهو مثال :

```
http://api.search.yahoo.com/WebSearchService/V1/webSearch?appid=Plagiar
ismDetector&query=plagiarism&results=2
```

حيث نقوم هنا بتمرير مفتاح التطبيق (PlagiarismDetector)- الذي تم طلبه من شركة Yahoo - وجملة البحث(plagiarism) ونقوم أيضا بتحديد عدد النتائج المراد جلبها وهي 2.

تعود النتائج مهيكلة (XML) حيث يكون أحد اللواحق المعادة (opening tag) :

- جميع نتائج البحث المتاحة (`totalResultsAvailable="3610652"`).
- جميع النتائج المعادة (`totalResultsReturned="2"`).
- موقع أول نتيجة (`firstResultPosition="1"`).

ومن ثم النتائج حيث تحتوي على عنوان الموقع المطلوب وبعض من محتوياته.
ومن خلال معالجة نتائج البحث نستطيع الحصول على عنوان الموقع المطلوب .

4.2.3.2 واجهة بحث Google البرمجية (Google AJAX Search API):

وهي مكتبة جافا سكريبت Javascript لاستخدام البحث عن طريق Google في صفحات الويب وتطبيقاته. ومن أجل البيئات الأخرى (التي لا تستخدم جافاسكريبت) تتيح Google واجهة تستخدم آلية REST وتعيد نتائج مهيكلة حسب JSON (معيار من أجل تنسيق النتائج) .

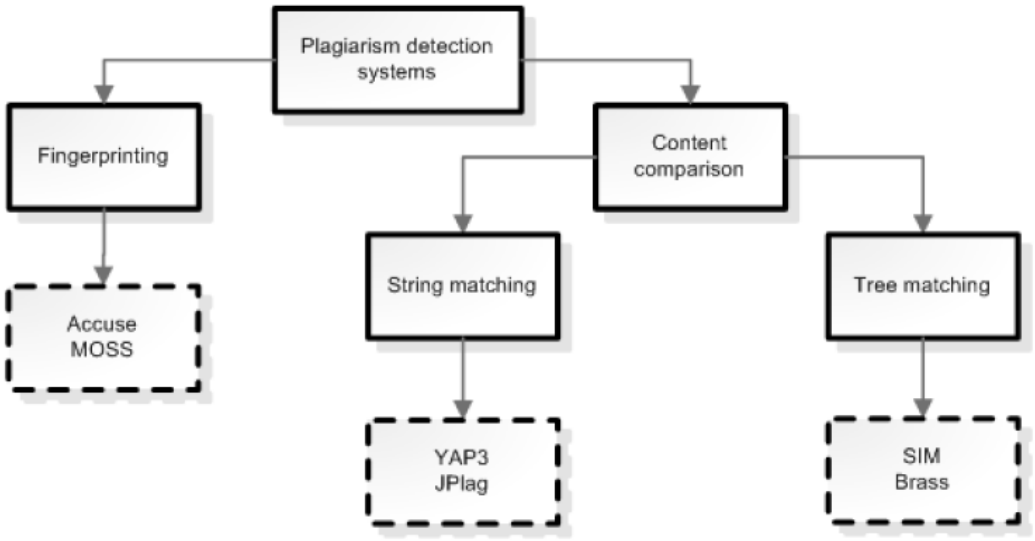
تم استخدام آلية REST مع JSON في المشروع نظرا لأن الأولى لا تسمح بالبحث التلقائي.

وتستخدم نفس الية Yahoo حسب نبني استعلام البحث باستخدام قواعد خاصة ثم نرسل هذا الاستعلام إلى مخدم البحث الخاص بـ Google ونقوم بمعالجة النتائج.

وأيضا هنا نحتاج إلى مفتاح (Key) من أجل السماح بعمليات البحث . هذا المفتاح يتم طلبه من شركة Google.

4.2.4 المقارن Comparer:

مهمة المقارن هي مقارنة نصين و تحديد التشابه بينهما وفقاً لخوارزمية ما. تصنف خوارزميات المقارنة لنوعيين رئيسيين موضحيين بالشكل التالي:



في النظام تم تطبيق خوارزميتي **Winnowing** و **LCS** و فيما يلي شرح مفّصل لهما:

4.2.4.1 خوارزمية Winnowing:

- خوارزميات Fingerprint:

إن مقارنة ملفات النصية يمكن أن يتم بشكل مباشر وفعال جداً باستخدام الخوارزميات التقليدية في حال كان المراد من المقارنة تحديد فيما إذا كانت الملفات متشابهة تماماً أم لا أما إذا كان المراد من المقارنة تحديد فيما إذا كان جزء ما من ملف معين مقتبس من ملف آخر فإن هذه الخوارزميات تصبح عاجزة تماماً ومستهلكة للوقت.

وهنا تأتي خوارزميات البصمة (Fingerprinting Algorithms) بالحل والتي تعمل على مقابلة ملف ذو حجم كبير بسلسلة رقمية ذات حجم أصغر بكثير من حجم الملف، تعرف ببصمة الملف حيث يفترض أن تعرِّف هذا الملف بشكل وحيد كما تعرِّف بصمة الإنسان صاحبها بشكل وحيد.[12] .

تستخدم هذه الخوارزمية في عمليات مقارنة المعطيات أو إرسالها عبر الشبكات الحاسوبية بحيث يتم تجنب الحجوم الكبيرة من المعطيات.[12] .

- خوارزمية Winnowing:[13] :

خوارزمية لاختيار بصمة للملفات النصية وبالتالي يمكن اعتبارها إحدى خوارزميات الـ Fingerprinting، تعمل هذه الخوارزمية على تقسيم الملف النصي إلى أجزاء متساوية الطول وطول كل منها k وعناصر هذه الأجزاء متتالية مباشرة في الملف الأصلي وهذا ما يعرف بـ k-grams. إن عدد هذه الأجزاء مساوٍ لعدد محارف النص منقوصاً منه طول السلسلة الجزئية حيث يتم أخذ أول k محرف كأول سلسلة جزئية ثم تتم إزاحة هذه السلسلة بمقدار محرف واحد لتتشكل لدينا سلسلة جديدة وهكذا حتى آخر محرف. ثم يتم ترميز كل مجموعة من المجموعات الجزئية برقم (hash) ويجب أن يميز هذا الرقم

مجموعته الجزئية بشكل وحيد، ثم يتم اختيار مجموعة جزئية من الأرقام (الرموز) الناتجة عن ترميز جميع المجموعات الجزئية لتكون هي بصمة الملف.

هناك العديد من الخوارزميات التي تقوم بما سبق لكن تختلف جميعها في طريقة اختيار المجموعة الجزئية الأخيرة، لذا فإننا سنسلط الضوء هنا على الخطوة الأخيرة من الخوارزمية (سبب الاختلاف).

تقوم هذه الخوارزمية بتعريف ما يسمى بالنافذة وهي قائمة من w عنصر تضع فيها في البداية أول w عنصر من قائمة الرموز الناتجة عن ترميز المجموعات الجزئية (k-grams)، ثم تختار منهم أصغر قيمة لتكون ضمن بصمة الملف النهائية وفي حال كانت هناك أكثر من قيمة مساوية لأصغر قيمة نختار القيمة في أقصى اليمين, ثم يتم إزاحة النافذة بمقدار عنصر واحد باتجاه نهاية قائمة الرموز السابقة (أي يتم حذف أول عنصر من الـ w عنصر السابقة وإزاحة باقي العناصر بحيث ينتقل العنصر من المكان i إلى المكان i+1 ثم يتم إضافة العنصر الجديد في آخر النافذة)، ونكرر العملية السابقة حتى انتهاء سلسلة الرموز.

وفي حال كانت أصغر قيمة في النافذة الحالية وحيدة ومشابهة لآخر قيمة تم اختيارها لتكون ضمن بصمة الملف النهائية يتم تجاهل هذه القيمة وتحريك النافذة مباشرة، ومن هنا يتم حذف بعض الرموز من بصمة الملف وبالتالي تصغير حجمها قدر المستطاع.

إن هذه الطريقة تعتمد على ما لوحظ تجريبياً بأنه غالباً ما تبقى القيمة الصغرى في النافذة الحالية هي نفسها في النافذة اللاحقة وعليه سيتم التوفير كثيراً في حجم البصمة المختارة للملف النصي، إلا أن ذلك يستدعي الاحتفاظ بموقع كل رمز بالإضافة لقيمته لنستطيع فيما بعد تحديد مواضع التشابه بين الملفات المقارنة.

بعد إيجاد بصمة لكل ملف وأصغر بكثير من حجم الملف يمكن الآن تطبيق إحدى خوارزميات المقارنة على بصمات الملفات كأن نطبق خوارزمية LCS والموضحة في هذا التقرير.

مثال:

ليكن لدينا النص التالي:

A do run run run, a do run run

1 - نحذف الفراغات والمحارف الخاصة (White Spaces).

Adorunrunrunadorunrun

2 - تقسيم النص إلى أجزاء متساوية وبطول 5 (5-grams).

adoru dorun orunr runru unrun nrunr runru unrun nruna runad unado nador adoru dorun orunr runru unrun

3 - ترميز الأجزاء السابقة (Hashing).

77 72 42 17 98 50 17 98 8 88 67 39 77 72 42 17 98

4 - تحريك نافذة من 4 عناصر على مجموعة الرموز السابقة.

(77, 74, 42, **17**) (74, 42, 17, 98) (42, 17, 98, 50) (17, 98, 50, **17**) (98, 50, 17, 98) (50, 17, 98, **8**) (17, 98, 8, 88) (98, 8, 88, 67) (8, 88, 67, 39) (88, 67, **39**, 77) (67, 39, 77, 74) (39, 77, 74, 42) (77, 74, 42, **17**) (74, 42, 17, 98)

5 - البصمة المختارة من خوارزمية الـ Winnowing:

17 17 8 39 17

4.2.4.2 خوارزمية السلسلة المشتركة الأطول (LCS):

خوارزمية لإيجاد أطول سلسلة جزئية مشتركة بين مجموعة من السلاسل (عادة تتألف هذه المجموعة من سلسلتين). وتحتل مكانة جيدة ضمن الخوارزميات المستخدمة لمقارنة الملفات لإيجاد درجة التشابه أو الاختلاف بينها. [14].

إن السلسلة الجزئية الناتجة هي سلسلة موجودة في كلا السلستين وعناصرها مرتبة بنفس ترتيب ورودها في كل منهما وليس بالضرورة أن تكون عناصر في السلسلتين متتالية. [15].

مثال: السلسلة المشتركة الأطول (LCS) للسلسلتين (ABC) و(ACB) هي أي من السلسلتين التاليتين (AB) و(AC). وبالتالي السلسلة المشتركة الأطول لمجموعة من السلاسل هي ليست سلسلة وحيدة.

- حل الخوارزمية من أجل سلسلتين:

إن مسألة السلسة المشتركة الأطول تتصف بأنها تمتلك ما يعرف بـ (Optimal Substructure)، إذ يمكن تقسيمها إلى مسائل أصغر وكل منها قابل للتقسيم لمسائل أصغر وهكذا إلى الوصول إلى مسألة ذات حل واضح وبسيط. كما أن هذه المسألة تمتلك أيضاً ما يعرف بـ (Overlapping Sub Problems)، إذ أن حل كل مسألة من المسائل الجزئية يعتمد على مجموعة الحلول للمسائل الجزئية من هذه المسألة. إن المسائل ذات الخواص السابقة (Optimal Substructure and Overlapping Sub Problems) يمكن حلها بواسطة البرمجة الديناميكية، حيث يتم بناء حل مسألة ما بدءاً من حلول مسائلها الجزئية. [14].

وبالتالي لحل هذه المسألة يجب الاحتفاظ بحلول مستوى معين من المسائل الجزئية بحيث يمكن الاعتماد عليها في حل المسائل الجزئية من المستوى الأعلى، وهذا ما يعرف بـ (Memoization) [16].

لحل المسألة السابقة من أجل سلسلتين ذوات طولين عشوائيين X, Y يمكن تحقيق الخطوتين التاليتين:

1 - في حال انتهاء كل من السلسلتين بنفس العنصر يتم حذف هذا العنصر من السلسلتين وإيجاد السلسلة المشتركة الأطول (LCS) للسلسلتين بعد حذف العنصر الأخير منهما، ثم دمج العنصر المحذوف في آخر السلسلة الناتجة.
2 - في حال عدم انتهاء كل من السلسلتين بنفس العنصر يتم حذف العنصر الأخير من السلسلة الأولى وإيجاد السلسلة المشتركة الأطول (LCS) لكل من السلسلتين الأولى بعد حذف العنصر الأخير والثانية كما هي، كما يتم حذف العنصر الأخير من السلسلة الثانية وإيجاد السلسلة المشتركة الأطول (LCS) لكل من السلسلتين الأولى كما هي و الثانية بعد حذف العنصر الأخير، ومنه تكون السلسلة المشتركة الأطول (LCS) للسلسلتين الأصليتين هي السلسة الأطول بين السلسلتين الناتجتين بعد عمليتي الحذف السابقتين.

يمكن التعبير عن الحل السابق كما يلي:

$$LCS(X_i, Y_i) = \begin{cases} \emptyset & if\ i = 0\ or\ j = 0 \\ Append(LCS(X_{i-1}, Y_{i-1}), X_i) & if\ X_i = Y_i \\ Longest\big(LCS(X_i, Y_{i-1}), LCS(X_{i-1}, Y_i)\big) & if\ X_i \neq Y_i \end{cases}$$

حيث:

- يعمل التابع Append على دمج العنصر Xi في آخر السلسلة:
  $LCS(X_{i-1}, Y_{i-1})$
- يقوم التابع Longest بإعادة السلسلة الأطول بين السلسلتين الممررتين له.

الملفات النصية:

تمثل الملفات النصية سلاسل طويلة من المحارف وبالتالي يمكن الاستفادة من خوارزمية السلسلة المشتركة الأطول لمقارنة هذه الملفات. ولكن في مقارنة الملفات النصية وخاصة ملفات اللغات الطبيعية كما هو الحال في النظام المعمول عليه فإن المهم هو تشابه العبارات (مجموعة من الجمل) وليس تشابه المحارف (عناصر هذه السلاسل (الملفات النصية)).

وبما أن خوارزمية السلسلة المشتركة الأطول تعطي النتائج بناء على مقارنة عناصر السلاسل وفي حالة هذا النظام المحارف فكان لابد من تعريف عتبة (Threshold) تمثل الطول الأقل الذي نستطيع عنده القول بأن تشابهاً بين الملفات قد حصل فعلاً، فمثلاً الجملة "ذهب فلان إلى الحديقة" لا تشابه "تناول فلان وجبته" بالرغم من أن نتيجة تطبيق الخوارزمية هي "فلان " وبوضع العتبة المذكورة سابقاً لقيمة كبيرة نسبياً "أكبر من 10 مثلاً" نعدل نتيجة الخوارزمية لتصبح نتيجة تطبيقها على المثال المعطى سلسلة فارغة.

إن العتبة الموضحة سابقاً حلت مشكلة لتظهر مع طريقة تحقيق الخوارزمية مشكلة أخرى نوضحها في المثال التالي:

لتكن لدينا الجملتين "درس فلان من الساعة الرابعة إلى الساعة الثامنة ثم تناول عشاءه وأوى إلى فراشه في تمام الساعة الثامنة والنصف" و "درس فلان من الساعة الرابعة إلى الساعة الثامنة" وهنا طبعاً فإن الجملة الثانية منسوخة كاملة من بداية الأولى. إلا أن تطبيق خوارزمية السلسلة المشتركة الأطول عليهما وبوضع العتبة على القيمة 15 سيؤدي إلى الناتج التالي " درس فلان من الساعة الرابعة " فقط وذلك لأن الخوارزمية ستطابق العبارتين كما يلي:

"درس فلان من الساعة الرابعة إلى الساعة الثامنة ثم تناول عشاءه وأوى إلى فراشه في تمام الساعة الثامنة والنصف"

"درس فلان من الساعة الرابعة إلى الساعة الثامنة"

وبالتالي كل من التطابقين "إلى" و"الساعة الثامنة" سيتم إهماله لأن عدد عناصره أقل من العتبة المحددة.

ولحل هذه المشكلة إما أن نلغي العتبة وهذا أمر مستحيل في اللغات الطبيعية إذ إن هناك كلمات تتكرر كثيراً في معظم الجمل كأحرف الجر وهذا ماسيؤدي بدوره إلى ضجيج في النتائج، أو نلجأ إلى ترميم النتائج يدوياً بأن نتأكد من أن المحرفين التاليين للنتيجة في كل من النصين المقارنين ليسا متساويين وفي حال كانا متساويين نضيف إلى النتيجة المحرف أحدهما ونفحص المحرفين التاليين لهما إلى أن نصل إلى محرفين غير متساويين وهي الطريقة المستخدمة في هذا النظام.

الفصل الخامس:

مراحل عمل النظام الرئيسية

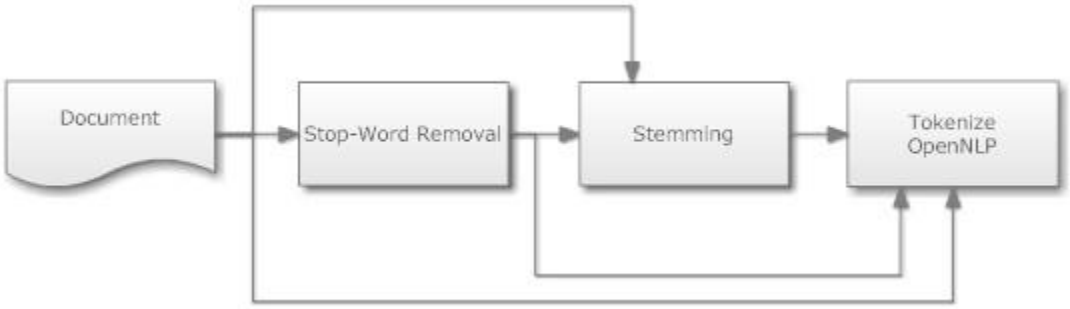
5

5. مراحل عمل النظام الرئيسية:

سنشرح مراحل العمل التالية:

1. تقطيع المستند *Tokenize*
2. ترتيب الـ Tokens الناتجة عن العملية السابقة *Ranking*
3. البحث باستخدام الانترنت للحصول على مصادر للمقارنة *Search for sources*
4. المقارنة *Comparison*
5. اظهار النتائج *View results*

5.1 تقطيع المستند Tokenize:



يكون في البداية دخل النظام عبارة عن نص (Plain text) مطلوب معالجته، لكن لكي نكون قادرين على تحديد الجزء المنتحل منه يجب أولاً أن نقسّمه لعدد من الجمل (Sentences) و تكون كل جملة بدورها مقطّعة لكلمات.

لماذا نستخدم Segment-based system؟

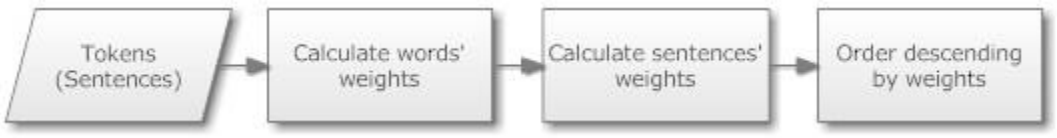
لأنه غالباً عندما يقوم الطالب بنسخ نص من مكان ما لا يمضي وقتاً طويلاً في مراجعته و إعادة كتابة جميع الجمل و بالتالي في حال تقطيع المستند إلى جمل يمكننا تحديد المسروقة منها بسهولة أكثر.

كما تتضمن هذه مرحلتين جزئيتين اختياريتين و هما:

i. حذف الكلمات غير المفيدة Stop-words
يتم تجاهل الكلمات غير المفيدة مثل أحرف الجر، أدوات الإشارة.

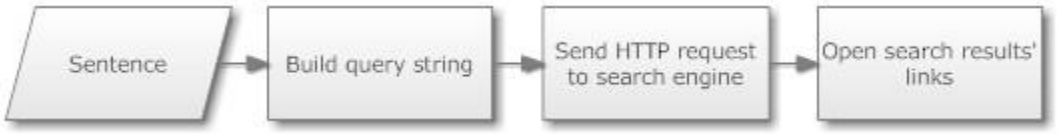
ii. إعادة الكلمات لأصلها Stemming
يتم حذف سوابق الكلمات و لواحقها. (Prefix, suffix) و ذلك باستخدام خوارزميتي Porter1 و Porter2

5.2 ترتيب الجمل Rank:



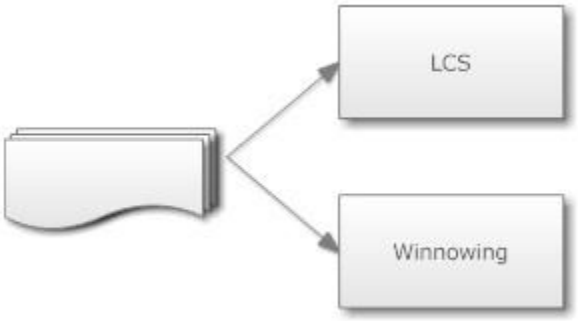
في هذه المرحلة يقوم النظام أولاً بحساب أوزان لكلمات النص، و من ثم اعتماداً على تلك الأوزان يقوم بحساب أوزان جمل النص باستخدام أحد التوابع الموضّحة سابقاً. وأخيراً ترتيب الجمل تنازلياً حسب أوزانها.

5.3 البحث Search for sources:



يبحث النظام عن الجملة الأكثر أهميةً أولاً إذا لم يرد محرك البحث أي نتيجة يستخدم النظام الجملة التالية للبحث و هكذا. تتم عملية البحث ببناء استعلام عن جملة ما يرسل عن طريق طلب HTTP لمحرك البحث و من ثم يتم تحليل نتائج البحث التي يردها محرك البحث المستخدم.

5.4 المقارنة Comparison:



يقارن النظام النص المدخل مع نتائج البحث التي حصل عليها من محرك البحث و ذلك وفق الخوارزمية التي يحددها المستخدم LCS أو Winnowing.

5.5 إظهار النتائج View results:

يقوم النظام باظهار نتائج عن طريق تلوين الكلمات المتشابهه و ذكر نسبة التشابه بين النص المدخل و نتائج البحث.

الفصل السادس:

التصميم و التحقيق

6

6. التصميم و التحقيق

6.1 التصميم:

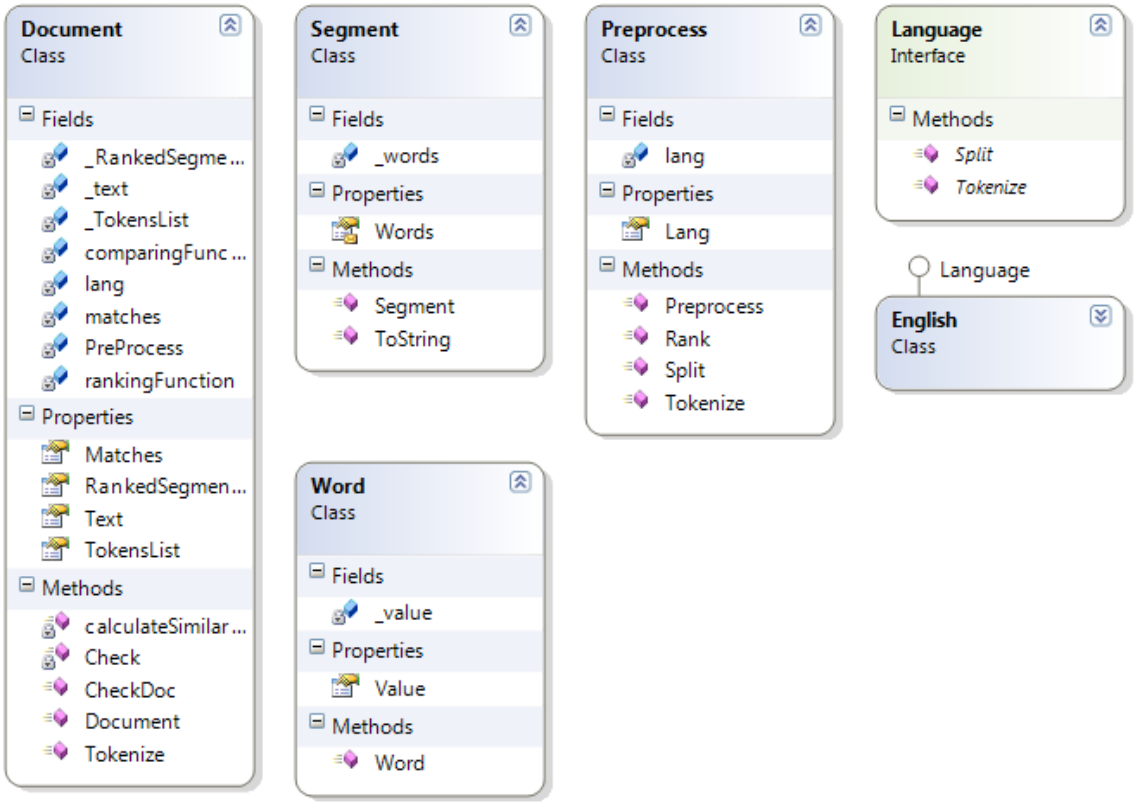
بالنسبة للأجزاء الدخلية للنظام سنقوم بتوضيح مخططات الصفوف الرئيسية فقط أما الأجزاء الخارجية سنكتفي فقط بذكر أسمائها و مهمة كل منها.

6.1.1 الأجزاء الخارجيّة:

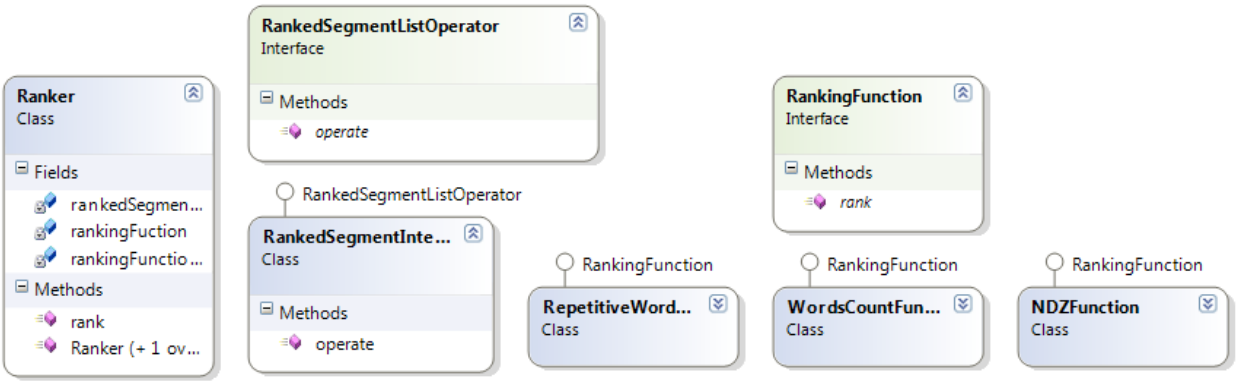
- OpenNLP: المحلل اللغوي.
- iTextSharp: مكتبة مجانية لقراءة الملفات من نمط PDF
- Majestic12: مكتبة مجانية و مفتوحة المصدر لقراءة صفحات HTML.

6.1.2 الأجزاء الداخلية:

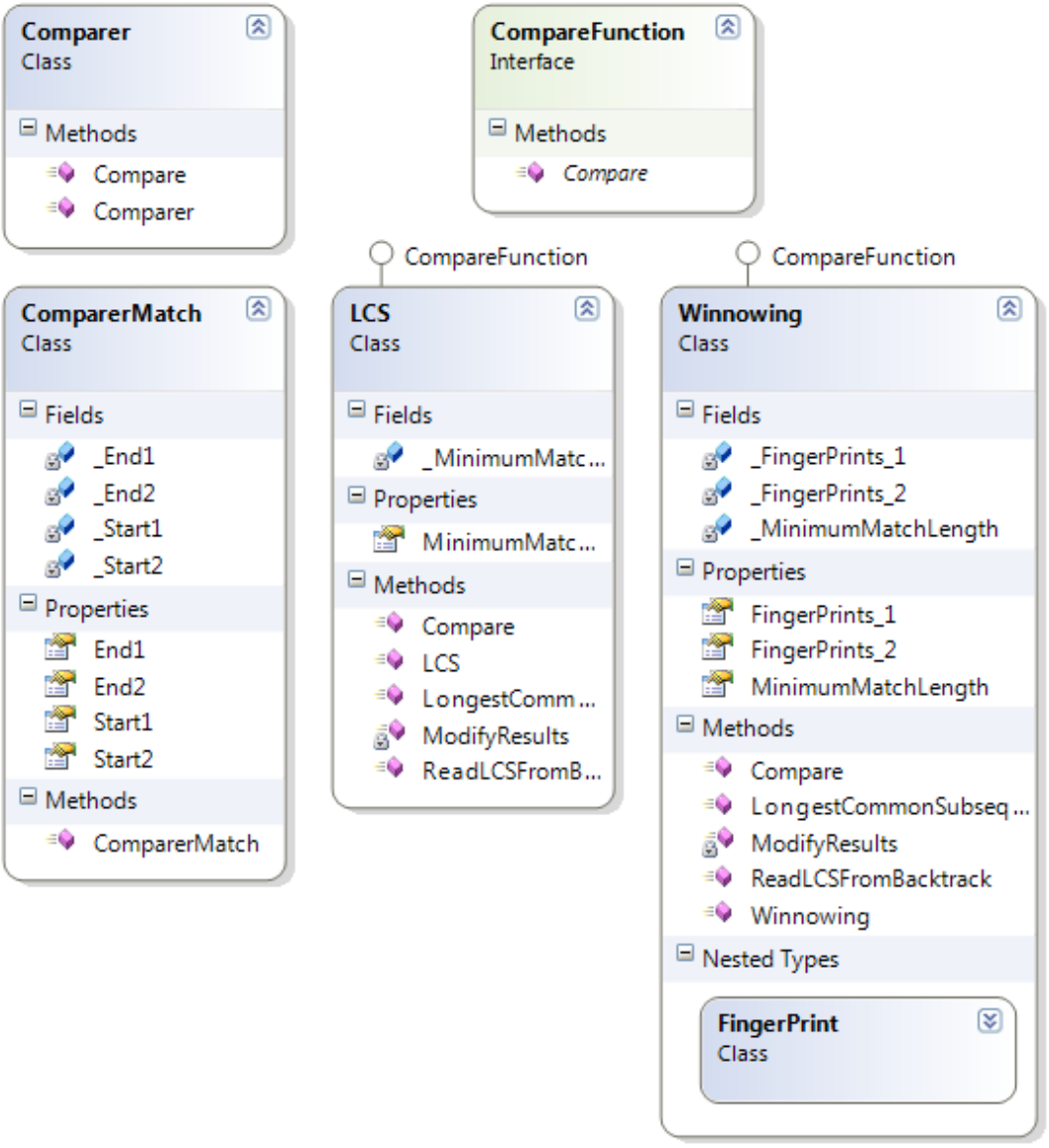
- **مخطط صفوف تمثيل مستند ما في النظام:**



- مخطط صفوف المرتّب Ranker:



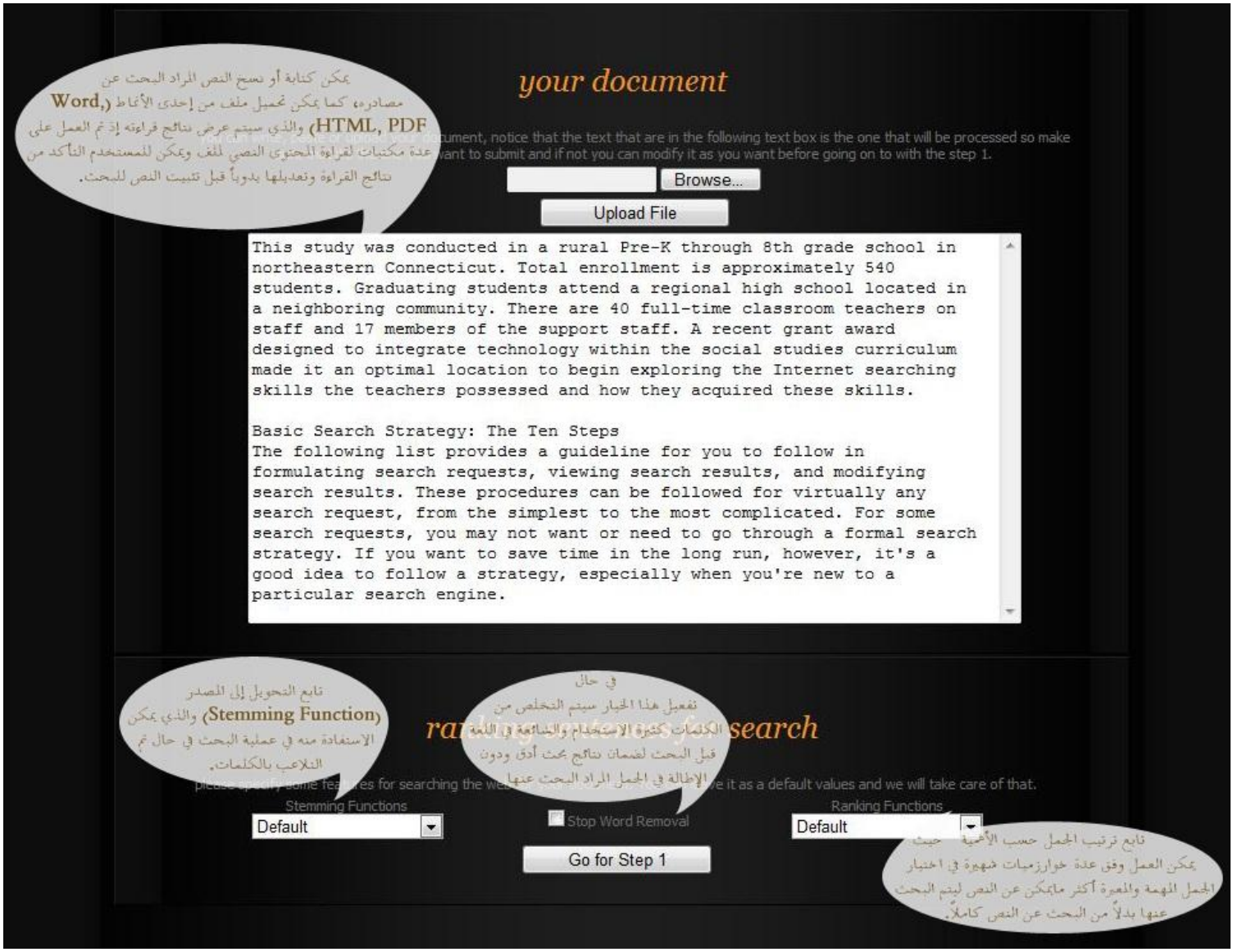
- مخطط صفوف المقارن Comparer:



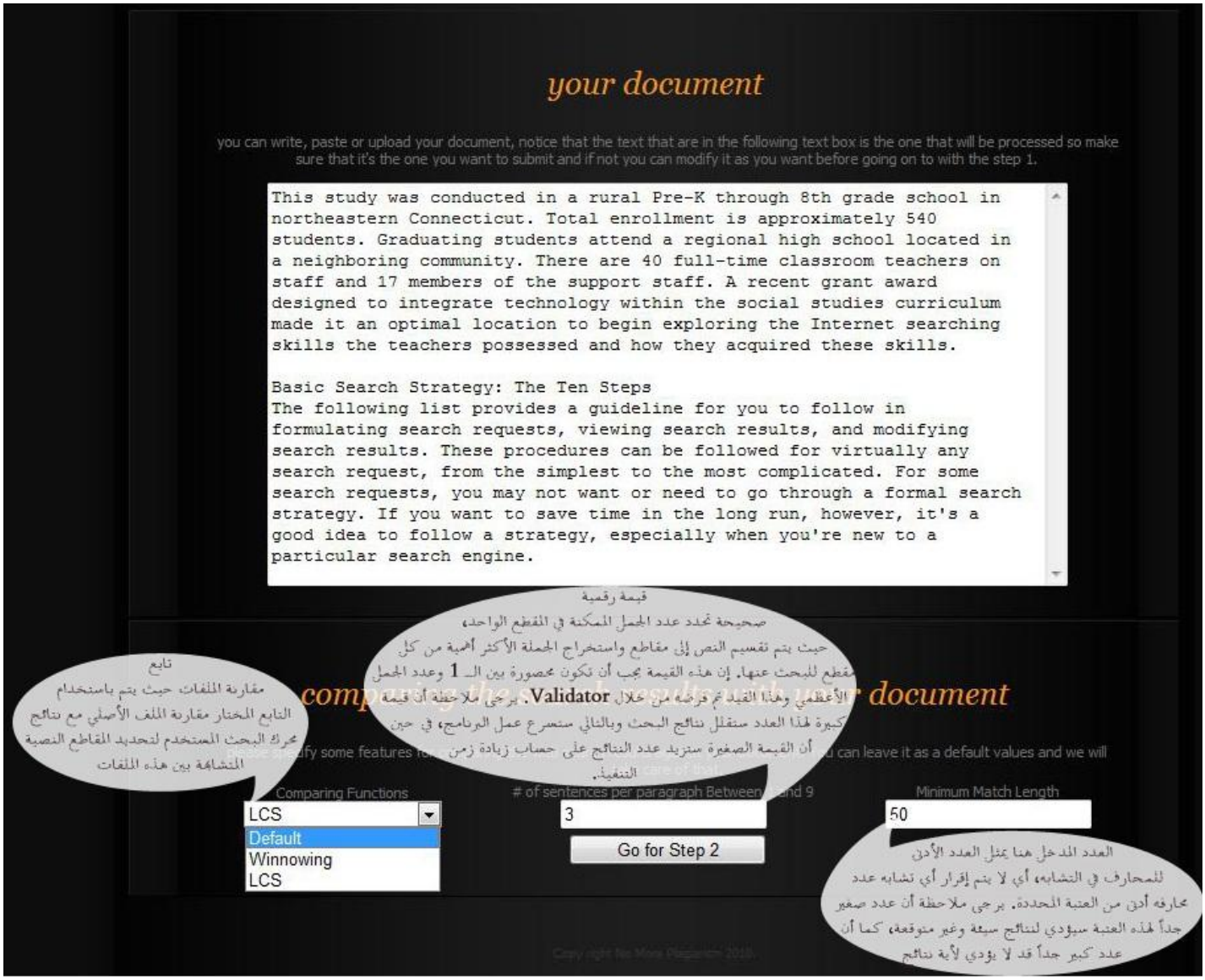
6.2 التحقيق:

في مايلي توضيح بعض واجهات التخاطب المستخدمة:

- الواجهة الرئيسية لتقديم ملف أو نص للمعالجة



- الواجهة الرئيسية لاختيار توابع المقارنة و تحديد عدد الجمل في الفقرة.



- الواجهة الرئيسية لعرض النتائج

الفصل السابع:

المشاكل

7

7. المشاكل

7.1مشاكل Google API:

نظرا لأن شركة Google لا تسمح باستخدام محرك بحثها من أجل البحث التلقائي (Automated search) فتمّ استخدامه محلياً. حيث أنها لا تمنع استخدامه من أجل عنوان IP ديناميكي (Dynamic IP address).

المشكلة الأساسية كانت عندما يصبح النظام على المخدم المضيف ويأخذ عنوان IP ثابت (Static IP address) تحجب شركة Google الخدمة عن الموقع السمتضيف للنظام و لذلك اضطررنا للعمل على النظام محلياً (Local host).

7.2مشاكل BOSS API :

وهي خدمة مقدمة من شركة Yahoo وهي اختصار لـ(*Build your Own Search Service*) أي استخدمها من أجل بناء خدمة البحث المرادة.

لم تعطي مشاكل مع عنوان الـ IP (مثل Google) سواء كان ثابت أم ديناميكي وإنما كانت نتائج البحث مختلفة وفي بعض الأحيان يعطي محرك بحث Google نتائج لا يعطيها محرك البحث التابع لـYahoo. وهذا موضح في الاختبارات.

الفصل الثامن:

الاختبارات

8

8. الاختبارات

8.1اختبارات توابع انتقاء الجمل الأكثر أهمية:

تم إجراء هذه الاختبارات مجموعة من الملفات المرفقة. وتم وضع اسم الملف بجانب كل اختبار وذلك لكبر حجم الملف واستحالة وضعه في التقرير.

تهدف الاختبارات إلى توضيح الفرق بين التوابع وفعاليتها في تمثيل الملف الأصلي المقتبسة منه وذلك باستخدام محرك البحث Google.

الاختبار الأول :

الملف: GameEngine

التابع	الجملة	ترتيب الملف الأصلي عند البحث في بواسطة Google
Word Count Function	[3] Features A rendered image can be understood in terms of a number of visible features [4] . . . at rates of approximately 20 to 120 frames per second .	لم تظهر
Repetitive Word Function	A game engine is a software system designed for the creation and development of video games.	النتيجة الأولى
NDZ Function	A game engine is a software system designed for the creation and development of video games.	النتيجة الأولى

الاختبار الثاني :

الملف: On Automatic Plagiarism Detection

التابع	الجملة	ترتيب الملف الأصلي عند البحث في بواسطة Google
Word Count Function	However , as we describe in the following section , the word-level n-grams comparison is not carried out considering sentences or entire documents . . . We must consider that plagiarised text fragments use to appear mixed and modified	النتيجة الأولى
Repetitive Word Function	On Automatic Plagiarism Detection Based on n-Grams Comparison Abstract .	النتيجة الأولى
NDZ Function	When automatic plagiarism detection is carried out considering a reference corpus , a suspicious text is compared to a set of original documents in order to relate the plagiarised text fragments to	النتيجة الأولى

	their potential source .	

الاختبار الثالث :

الملف: Shared Information and Program Plagiarism Detection

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	2.1 Attribute Counting Systems The earliest attribute-counting-metric system [14] used Halsted 's software science metrics to measure the level of similarity between program pairs . . . over all distinct types .	**Word Count Function**
النتيجة الأولى	Shared Information and Program Plagiarism Detection Abstract A fundamental question in information theory and in computer science is how to measure similarity or the amount of shared information between two sequences .	**Repetitive Word Function**
النتيجة الأولى	Shared Information and Program Plagiarism Detection Abstract A fundamental question in information theory and in computer science is how to measure similarity or the amount of shared information between two sequences .	**NDZ Function**

الاختبار الرابع :

الملف : A Plagiarism Detection Tool

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
لم تظهر	We have shown that this tool works as well as other copy detection tools. . . . . Introduction A copy detection software is a piece of software able to identify equal parts between two or more files .	**Word Count Function**
لم تظهر	A Plagiarism Detection Tool Abstract Plagiarism in student programming assignments is a possibility which needs to be taken into account when a group of students are working on the same project	**Repetitive Word Function**
لم تظهر	A Plagiarism Detection Tool Abstract Plagiarism in student programming assignments is a possibility which needs to be taken into account when a group of students are working on the same project	**NDZ Function**

الاختبار الخامس :

الملف : A Web-Enabled Plagiarism Detection Tool

التابع	الجملة	ترتيب الملف الأصلي عند البحث في بواسطة Google
Word Count Function	Returning to higher education , universities around the world are ramping up their efforts against plagiarism . . . . but the issue of detection has not received enough attention ; as we said ,we have discovered only a handful of cases at our campus .	النتيجة الأولى
Repetitive Word Function	A Web-Enabled Plagiarism Detection Tool This material is presented to ensure timely dissemination of scholarly and technical work .	النتيجة الأولى
NDZ Function	A Web-Enabled Plagiarism Detection Tool This material is presented to ensure timely dissemination of scholarly and technical work .	النتيجة الأولى

الاختبار السادس :

الملف : Extending Web Search for Online Plagiarism Detection

التابع	الجملة	ترتيب الملف الأصلي عند البحث في بواسطة Google
Word Count Function	Consequently , several prior works [3] [11] [12] [17] [18] are proposed to relieve the impact of this problem on search engines and large databases Several types of plagiarism can be enumerated [6] to reflect the degree or the seriousness of the plagiarism problem .	النتيجة الأولى
Repetitive Word Function	Extending Web Search for Online Plagiarism Detection Abstract As information technologies advance , the data amount gathered on the Internet increases at an incredible rapid speed .	النتيجة الأولى
NDZ Function	To solve the data overloading problem , people commonly use web search engines to find what they need .	النتيجة الخامسة

الاختبار السابع :

الملف : Old and new challenges in automatic plagiarism detection

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	Plagiarism detection The aim of this paper is to present plagiarism detection as a problem to be solved … or guidance for writers on how to prevent themselves unintentionally plagiarising their sources .	**Word Count Function**
النتيجة الأولى	Old and new challenges in automatic plagiarism detection Automatic methods of measuring similarity between program code and natural language text pairs have been used for many years to assist humans in detecting plagiarism .	**Repetitive Word Function**
النتيجة الأولى	Old and new challenges in automatic plagiarism detection Automatic methods of measuring similarity between program code and natural language text pairs have been used for many years to assist humans in detecting plagiarism .	**NDZ Function**

الاختبار الثامن :

الملف : Plagiarism detection using software tools

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	In the area of computer science and related degrees this issue presents some particular features such as … Sometimes , due to the number of projects or the nature of the proposed activities , reviewing work is done in a distributed manner by several instructors .	**Word Count Function**
النتيجة الأولى	Plagiarism detection using software tools : a study in a Computer Science degree Keywords Plagiarism prevention and detection , e-learning , e-evaluation .	**Repetitive Word Function**
النتيجة الأولى	Plagiarism presents particular features in Computer Science and related degrees , such as … sharing of knowledge has to be promoted among students.	**NDZ Function**

الاختبار التاسع :

الملف : SNITCH A Software Tool for Detecting

التابع	الجملة	ترتيب الملف الأصلي عند البحث في بواسطة Google
Word Count Function	In general , the algorithm uses the following steps : □ Open a document □ Analyze the document … and other pertinent statistics .	النتيجة الأولى
Repetitive Word Function	SNITCH : A Software Tool for Detecting Cut and Paste Plagiarism ABSTRACT Plagiarism of material from the Internet is a widespread and growing problem .	النتيجة الأولى
NDZ Function	Computer science students , and those in other science and engineering courses , can sometimes get away with a “cut and paste” approach to assembling a paper in part because the expected style of technical writing is less expositional than in liberal arts courses .	النتيجة الأولى

8.2اختبارات واجهات محركات البحث:

تم تنفيذ هذا النوع من الاختبارات لكي نستطيع تحديد أي محرك بحث أفضل بالنسبة للنظام. يتم الاختبار بارسال طلب HTTP عن طريق واجهة التطبيق البرمجية لمحرك البحث. يتضمن هذا الطلب جملة أُخذت من موقع أو مقالة ما.

الاختبار الأول :

Several researches developed optimized ontology-based semantic (OBSC) framework for English content. The methodology used in these approaches could not be used for Arabic content due to the complexity of the syntax, semantics and ontology of the Arabic language.

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	لم تظهر
Google	1

الاختبار الثاني :

On Automatic Plagiarism Detection Based on n-Grams Comparison Abstract .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

الاختبار الثالث :

Shared Information and Program Plagiarism Detection Abstract A fundamental question in information theory and in computer science is how to measure similarity or the amount of shared information between two sequences .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

الاختبار الرابع :

SNITCH : A Software Tool for Detecting Cut and Paste Plagiarism ABSTRACT Plagiarism of material from the Internet is a widespread and growing problem .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

الاختبار الخامس :

A Web-Enabled Plagiarism Detection Tool This material is presented to ensure timely dissemination of scholarly and technical work .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

الاختبار السادس :

Extending Web Search for Online Plagiarism Detection Abstract As information technologies advance , the data amount gathered on the Internet increases at an incredible rapid speed .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

الاختبار السابع :

Old and new challenges in automatic plagiarism detection Automatic methods of measuring similarity between program code and natural language text pairs have been used for many years to assist humans in detecting plagiarism .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

الاختبار الثامن :

Plagiarism detection using software tools : a study in a Computer Science degree Keywords Plagiarism prevention and detection , e-learning , e-evaluation .

واجهة البحث المستخدمة	ترتيب ظهور المقال التابعة له في نتائج البحث
Yahoo	1
Google	1

النتائج:

- محرك بحث Google يستطيع إيجاد مقالات و مواقع منشورة حديثاً، أي صفحات و ملفات جديدة نوعاً ما، أكثر من محرك بحث Yahoo. و يتضح ذلك من الاختبار الأول فالجملة المستخدمة في البحث هي جملة من مقال نشر منذ شهرين فقط.

- خلال الاختبارات تبيّن أن استعلامات محرك البحث Google عن طريق واجهة التطبيق البرمجية API لا تعطي نتائج مطابقة تماماً للنتائج التي يعطيها محرك بالبحث عند قيام مستخدم ما بالبحث عن طريق موقع محرك البحث. مثال ذلك الجملة التالية: *"The preceding examples were based on fixed-length codes , such as 12-bit numbers encoding values between 1 and 4,000".* عند البحث عن هذه الجملة عن طريق Google API تكون النتيجة الأولى هي www.cs.cmu.edu/~dst/Tutorials/Info-Theory بينما عندما نبحث عن طريق موقع Google تكون النتيجة الأولى هي profile.iiita.ac.in/pkmaurya_b03/from%20mail/dc1.doc

- بشكل عام، محرك بحث Google يعطي نتائج أفضل من Yahoo لذلك تم اختياره كمحرك بحث افتراضي للنظام.

8.3اختبارات خوارزميات المقارنة:

استخدمنا في المقارنة المصطلحات التالية للتعبير عن سير البرنامج :

الإيجابية الخاطئة : توجد عندما يكون النص غير منتحل(سلبية) ويظهر البرنامج أن النص منتحل (نتيجة خاطئة)
وعبرنا عن هذه الخاصة باللون الأخضر.
السلبية الخاطئة : توجد عندما يكون النص منتحل(إيجابية) ويظهر البرنامج أن النص غير منتحل (نتيجة خاطئة).
وعبرنا عن هذه الخاصة باللون الأحمر.
اللون الأصفر للتعبير عن الجزء المشابه بين النصين.

الاختبار الأول :

الخوارزمية	عدد حروف النص الأصلي	عدد حروف النص المقتبس	زمن المقارنة (ميلي ثانية)	نسبة الإيجابية الخاطئة	نسبة السلبية الخاطئة
Winnowing	575	512	25	0.5 %	0 %

النص الأصلي	النص المقتبس
An earthquake (also known as a quake, tremor, temblor or seismic activity) is the result of a sudden release of energy in the Earth's crust that creates seismic waves. Earthquakes are measured with a seismometer; a device which also records is known as a *seismograph.* The moment magnitude (or the related and mostly obsolete Richter magnitude) of an earthquake is conventionally reported, with magnitude 3 or lower earthquakes being mostly imperceptible and magnitude 7 causing serious damage over large areas. Intensity of shaking is measured on the modified Mercalli scale.	An earthquake (also known as a quake, tremor, temblor or seismic activity) is the result of a sudden release of energy in the Earth's crust that creates seismic waves. There are three main types of fault that may cause an earthquake: normal, reverse (thrust) and strike-slip. Normal and reverse faulting are examples of dip-slip, where the displacement along the fault is in the direction of dip and movement on them involves a vertical component. Intensity of shaking is measured on the modified Mercalli scale.

الاختبار الثاني :

الخوارزمية	عدد حروف النص الأصلي	عدد حروف النص المقتبس	زمن المقارنة (ميلي ثانية)	نسبة الإيجابية الخاطئة	نسبة السلبية الخاطئة
LCS	575	512	30	0.5 %	0 %

النص الأصلي	النص المقتبس
An earthquake (also known as a quake, tremor, temblor or seismic activity) is the result of a sudden release of energy in the Earth's crust that creates seismic waves. Earthquakes are measured with a seismometer; a device which also records is known as a *seismograph.* The moment magnitude (or the related and mostly obsolete Richter magnitude) of an earthquake is conventionally reported, with magnitude 3 or lower earthquakes being mostly imperceptible and magnitude 7 causing serious damage over large areas. Intensity of shaking is measured on the modified Mercalli scale.	An earthquake (also known as a quake, tremor, temblor or seismic activity) is the result of a sudden release of energy in the Earth's crust that creates seismic waves. There are three main types of fault that may cause an earthquake: normal, reverse (thrust) and strike-slip. Normal and reverse faulting are examples of dip-slip, where the displacement along the fault is in the direction of dip and movement on them involves a vertical component. Intensity of shaking is measured on the modified Mercalli scale.

الاختبار الثالث:

الخوارزمية	عدد حروف النص الأصلي	عدد حروف النص المقتبس	زمن المقارنة (ميلي ثانية)	نسبة الإيجابية الخاطئة	نسبة السلبية الخاطئة
Winnowing	1524	1156	95	0.9 %	12 %

النص الأصلي	النص المقتبس
A tectonic earthquake begins by an initial rupture at a point on the fault surface, a process known as nucleation. The scale of the nucleation zone is uncertain, with some evidence, such as the rupture dimensions of the smallest earthquakes, suggesting that it is smaller than 100 m while other evidence, such as a slow component revealed by low-frequency spectra of some earthquakes, suggest that it is larger. The possibility that the nucleation involves some sort of preparation process is supported by the observation that about 40% of earthquakes are preceded by foreshocks. Once the rupture has initiated it begins to propagate along the fault surface. The mechanics of this process are poorly understood, partly because it is difficult to recreate the high sliding velocities in a laboratory. Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone. Sometimes a series of earthquakes occur in a sort of earthquake storm, where the earthquakes strike a fault in clusters, each triggered by the shaking or stress redistribution of the previous earthquakes. Similar to aftershocks but on adjacent segments of fault, these storms occur over the course of years, and with some of the later earthquakes as damaging as the early ones. Such a pattern was observed in the sequence of about a dozen earthquakes that struck the North Anatolian Fault in Turkey in the 20th century and has been inferred for older anomalous clusters the Middle East.	Most earthquakes form part of a sequence, related to each other in terms of location and time. Most earthquake clusters consist of small tremors which cause little to no damage, but there is a theory that earthquakes can recur in a regular pattern. The scale of the nucleation zone is uncertain, with some evidence. Earthquake swarms are sequences of earthquakes striking in a specific area within a short period of time. Once the rupture has initiated it begins to propagate along the fault surface. Earthquakes often occur in volcanic regions and are caused there, both by tectonic faults and the movement of magma in volcanoes. Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone. Earthquake swarms can serve as markers for the location of the flowing magma throughout the volcanoes. Similar to aftershocks but on adjacent segments of fault, these storms occur over the course of years, and with some of the later earthquakes as damaging as the early ones. These swarms can be recorded by seismometers and tiltmeters. about a dozen earthquakes that struck the North Anatolian Fault in Turkey.

الاختبار الرابع:

الخوارزمية	عدد حروف النص الأصلي	عدد حروف النص المقتبس	زمن المقارنة (ميلي ثانية)	نسبة الإيجابية الخاطئة	نسبة السلبية الخاطئة
LCS	1524	1156	165	3.0 %	12 %

النص الأصلي	النص المقتبس
A tectonic earthquake begins by an initial rupture at a point on the fault surface, a process known as nucleation. The scale of the nucleation zone is uncertain, with some evidence, such as the rupture dimensions of the smallest earthquakes, suggesting that it is smaller than 100 m while other evidence, such as a slow component revealed by low-frequency spectra of some earthquakes, suggest that it is larger. The possibility that the nucleation involves some sort of preparation process is supported by the observation that about 40% of earthquakes are preceded by foreshocks. Once the rupture has initiated it begins to propagate along the fault surface. The mechanics of this process are poorly understood, partly because it is difficult to recreate the high sliding velocities in a laboratory. Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone. Sometimes a series of earthquakes occur in a sort of earthquake storm, where the earthquakes strike a fault in clusters, each triggered by the shaking or stress redistribution of the previous earthquakes. Similar to aftershocks but on adjacent segments of fault, these storms occur over the course of years, and with some of the later earthquakes as damaging as the early ones. Such a pattern was observed in the sequence of about a dozen earthquakes that struck the North Anatolian Fault in Turkey in the 20th century and has been inferred.	Most earthquakes form part of a sequence, related to each other in terms of location and time. Most earthquake clusters consist of small tremors which cause little to no damage, but there is a theory that earthquakes can recur in a regular pattern. The scale of the nucleation zone is uncertain, with some evidence. Earthquake swarms are sequences of earthquakes striking in a specific area within a short period of time. Once the rupture has initiated it begins to propagate along the fault surface. Earthquakes often occur in volcanic regions and are caused there, both by tectonic faults and the movement of magma in volcanoes. Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone. Earthquake swarms can serve as markers for the location of the flowing magma throughout the volcanoes. Similar to aftershocks but on adjacent segments of fault, these storms occur over the course of years, and with some of the later earthquakes as damaging as the early ones. These swarms can be recorded by seismometers and tiltmeters. about a dozen earthquakes that struck the North Anatolian Fault in Turkey.

النتائج:

- خوارزمية Winnowing أسرع من LCS بشكل عام.
- خوارزمية LCS لا تصلح للتعامل مع الملفات كبيرة الحجم

8.4اختبارات النظام كاملاً:

الاختبار الأول:

عدد الجمل في الفقرة	نسبة السلبية الخاطئة	نسبة الإيجابية الخاطئة	الزمن (ثانية)	عدد حروف النص	خوارزمية المقارنة	محرك البحث	تابع انتقاء الجمل
14	9%	0%	119	1394	Winnowing	Google	NDZ

النص

When you are programming with threads, understanding the life cycle of thread is very valuable. While a thread is alive, it is in one of several states. By invoking start() method, it doesn't mean that the thread has access to CPU and start executing straight away. Several factors determine how it will proceed.

1. New state – After the creations of Thread instance the thread is in this state but before the start() method invocation. At this point, the thread is considered not alive.

2. Runnable (Ready-to-run) state – A thread start its life from Runnable state. A thread first enters runnable state after the invoking of start() method but a thread can return to this state after either running, waiting, sleeping or coming back from blocked state also. On this state a thread is waiting for a turn on the processor.

3. Running state – A thread is in running state that means the thread is currently executing. There are several ways to enter in Runnable state but there is only one way to enter in Running state: the scheduler select a thread from runnable pool.

4. Dead state – A thread can be considered dead when its run() method completes. If any thread comes on this state that means it cannot ever run again.

5. Blocked - A thread can enter in this state because of waiting the resources that are hold by another thread.

المواقع التي تم الاقتباس عنها :

http://www.roseindia.net/java/thread/life-cycle-of-threads.shtml

الاختبار الثاني :

تابع انتقاء الجمل	محرك البحث	خوارزمية المقارنة	عدد حروف النص	الزمن (ثانية)	نسبة الإيجابية الخاطئة	نسبة السلبية الخاطئة	عدد الجمل في الفقرة
NDZ	Google	Winnowing	1575	191	%0	%0	7

النص

Dynamic Programming (DP) generates all enumerations, or rather, cases of the smaller breakdown problems, leading towards the larger cases, and eventually it will lead towards the final enumeration .of size n. As in Fibonacci numbers, DP generated all Fibonacci numbers up to n
.Once you are given a problem, it is usually a good idea to check if DP is applicable to it
The second step to solving a problem using DP is to recognize the recursive relationship. The relationship maybe straightforward or even pointed out, or it maybe hidden and you have to find it. In any case, since you have already determined that it is indeed a DP problem, you should at least have a .pretty good idea of the relationship

I find Markdown to be a more readable and usable alternative to XHTML/CSS for formatting text, and I use it to format my articles at this Django-powered blog. When implementing syntax highlighting for code blocks within text, I searched for existing solutions and found many approaches that were too complicated and had shortcomings. After more research, I realized that syntax highlighting works out .of the box in Django if you have a recent version of Markdown

Here are the required steps to enable syntax highlighting in your Django application. First, install python-markdown version 2.0+ and python-pygments. Pygments is a syntax highlighter written in Python. Markdown 2.0+ has an extension system and comes with a syntax highlighting extension that uses Pygments. This extension is called CodeHilite. To use it, add the following to a Django template

المواقع التي تم الاقتباس عنها :

http://www.algorithmist.com/index.php/Dynamic_Programming
http://aymanh.com/syntax-highlighting-django-markdown-pygments

الاختبار الثالث :

عدد الجمل في الفقرة	نسبة السلبية الخاطئة	نسبة الإيجابية الصحيحة	الزمن (ثانية)	عدد حروف النص	خوارزمية المقارنة	محرك البحث	تابع انتقاء الجمل
4	0%	0.11%	529137.2649	3101	Winnowing	Google	NDZ

النص

I find Markdown to be a more readable and usable alternative to XHTML/CSS for formatting text, and I use it to format my articles at this Django-powered blog. When implementing syntax highlighting for code blocks within text, I searched for existing solutions and found many approaches that were too complicated and had shortcomings. After more research, I realized that syntax highlighting works out of the box in Django if you have a recent version of Markdown.

Maintenant, je crois que la profession qui je souhaite exercer est clair: Chercheur Chercheur, qu'est-ce que ça veux dire?

1- Un chercheur (féminine chercheuse) une personne dont le métier consiste à faire de la recherche.

2- Selon la définition de l'organisation de coopération et de développement économiques le chercheur est :

« Spécialiste travaillant à la conception ou à la création de connaissances, de produits, de procédés, de méthodes et de systèmes nouveaux et à la gestion des projets concernés »

Quels sont les diplômes nécessaires? Pour devenir chercheur vous devrez obtenir un doctorat, mais qu'est-ce que on doit faire pour obtenir le grade de docteur?

Premièrement, préparez une thèse et quand vous êtes prêt vous allez présenter votre travail devant un jury académique et selon votre travail ils décident est-ce que vous méritez le doctorat ou non….

Here are the required steps to enable syntax highlighting in your Django application. First, install python-markdown version 2.0+ and python-pygments. Pygments is a syntax highlighter written in Python. Markdown 2.0+ has an extension system and comes with a syntax highlighting extension that uses Pygments. This extension is called CodeHilite. To use it, add the following to a Django template.

Est-ce qu'il faut suivre des études à l'étranger? Pas nécessaire, mais, bien sûr il va être mieux si on les suit, puisque ici, en Syrie, il n'y pas beaucoup de laboratoire.

Les qualités nécessaires: Le chercheur doit être patient, compétent, persévérant et peut-être curieux.

Les avantages: Intéressant, pas traditionnel, chaque jour il y a quelque chose nouvelle.

Les inconvénients: La vie quotidienne d'un chercheur est une vie souvent étrange car il faut en moyenne 15 ans à un chercheur pour faire et valider une découverte valable, donc une recherche est trop longue à faire et il est possible que le chercheur subit un échec à la fin!

Dynamic Programming (DP) generates all enumerations, or rather, cases of the smaller breakdown problems, leading towards the larger cases, and eventually it will lead towards the final enumeration of size n. As in Fibonacci numbers, DP generated all Fibonacci numbers up to n.

Once you are given a problem, it is usually a good idea to check if DP is applicable to it.

The second step to solving a problem using DP is to recognize the recursive relationship. The relationship maybe straightforward or even pointed out, or it maybe hidden and you have to find it. In any case, since you have already determined that it is indeed a DP problem, you should at least have a pretty good idea of the relationship.

المواقع التي تم الاقتباس عنها :

http://aymanh.com/syntax-highlighting-django-markdown-pygments
http://fr.wikipedia.org/wiki/Chercheur
http://www.algorithmist.com/index.php/Dynamic_Programming

الفصل التاسع:

الخلاصة و الآفاق المستقبلية

9

9. الخلاصة و الآفاق المستقبلية:

9.1 الخلاصة:

ناقشنا في هذا التقرير نظاماً لكشف الانتحال في نصوص اللغات الطبيعية. تميّز النظام عن الأنظمة الأخرى بخاصية الاختيار التلقائي للجمل التي يجب البحث عنها على الانترنت حيث اعتمدنا على مبدأ توزين الكلمات. كم أتاح النظام للمستخدم المجال لكي يحدد عمق الاختبار و أقصر طول لسلسلة محارف يمكن أن تعتبر منتحلة، و تحديد خوارزمية المقارنة و معاملاتها.

أمّا مشاكل النظام فكانت الزمن الطويل جداً، و أحياناً الفشل، في معالجة الكتب و المقالات الطويلة جداً. نتطلع لحل هذه المشكلة في المستقبل عن طريق تغيير خوارزميات المقارنة.

و أخيراً، يمكن أن يكون هذا النظام مساعداً للمدرسين في تفحص وظائف طلابهم، و لناشري المقالات في تحديد ما تم اقتباسه من مقالاتهم. وهذا يتم كخدمة تقدم لهم في منازلهم عن طريق شبكة الانترنت فما عليهم سوى تحميل الوظيفة أو المقالة وانتظار النتائج.

9.2 الآفاق المستقبلية:

نتطلع في المستقبل لإضافة خيارات و تحسينات عديدة على النظام منها:

- دعم اللغة العربيّة بشكل أفضل و إضافة Stemmer خاص باللغة العربية على النظام.
- المقارنة حسب المعنى:
  - استخدام خوارزميات المقارنة التي تعتمد على المعنى التي تعطي نتائج دقيقة جداً لكنها تستهلك وقتاً أكثر من الخوارزميات الإحصائية. من هذه الخوارزميات: Running-Karp-Rabin Greedy-String-Tiling (RKR-GST) و FPDS
  - استبدال الأرقام في النصوص بشكلها الحرفي Number Replacement [7]
  - تعميم المفردات Word Generalization [7] فنستبدل كلمة قطة بكلمة حيوان أليف مثلاً.
- تطوير البحث وذلك بالاستفادة من عدة محركات بحث في نفس الوقت ومقاطعة نتائجهم.
- إضافة خوارزميات جديدة لاختيار الجمل المعبرة عن النص قد تكون أكثر دقة من الحالية.

ملحق آ:

المراجع

آ

محلق آ : المراجع

[1] C.J. Neill and G. Shanmuganathan. “A Web-Enabled Plagiarism Detection Tool,” *IEEE IT Professional*, Vol. 6, No. 5, September-October 2004. pp. 19-23.

[2] Sebastian Niezgoda and Thomas P. Way. "SNITCH: a Software Tool for Detecting Cut and Paste Plagiarism". *SIGCSE Technical Symposium (SIGCSE 2006)*, pages 51-55, March 2006

[3]Regent Court. “Old and new challenges in automatic plagiarism detection”. *University of Sheffield*

[4] Mozgovoy, M. (2007). "Enhancing computer-aided plagiarism detection". *Doctoral Thesis, University of Joensuu*, Department of Computer Science and Statistics

[5] Yi-Ting Liu, Heng-Rui Zhang, Tai-Wei Chen and Wei-Guang Teng. (2007). "Extending Web Search for Online Plagiarism Detection". *National Cheng Kung University, Taiwan*

[6] G. Salton and C. Buckley. “Term-Weighting Approaches in Automatic Text Retrieval,” *Information Processing & Management* 24(5): 513–523, 1988.

[7] Zdeněk Češka and Chris Fox. "The Influence of Text Pre-processing on Plagiarism Detection". *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria, pp. 55-59, September 2009.

[8] http://www.kuleuven.be/plagiarism/examples.html

[9] *Plagiarism detection*. Retrieved from Wikipedia.org: http://en.wikipedia.org/wiki/Plagiarism_detection

[10] D. McCabe. Levels of Cheating and Plagiarism Remain High. Center for Academic Integrity - Duke University, 2005. Website: http://www.academicintegrity.org.

[11] Daniela Chudaa and Pavol Navrata. (2010) “Support for checking plagiarism in e-learning”. *Slovak University of Technology*, Ilkovicova 3, 812 19 Bratislava, Slovakia.

[12]*Fingerprint (computing)*. Retrieved from Wikipedia.org: http://en.wikipedia.org/wiki/Fingerprint_(computing)

[13]S. Schleimer, D. S. Wilkerson, and A. Aiken. “Winnowing: Local algorithms for document fingerprinting”. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data 2003*, pages 76--85. *ACM Press*, 2003.

[14] *Longest common subsequence problem*. Retrieved from Wikipedia.org: http://en.wikipedia.org/wiki/Longest_common_subsequence_problem#cite_note-0

[15]*Longest Common Subsequence*. Retrieved from Algorithmist.com

http://www.algorithmist.com/index.php/Longest_Common_Subsequence

[16] *Memoization*. Retrieved from Wikipedia.org: http://en.wikipedia.org/wiki/Memoization

[17]*OpenNLP Home*. Retrieved from opennlp.sourceforge.net: http://opennlp.sourceforge.net/

[18] *OpenNLP Projects*. Retrieved from opennlp.sourceforge.net: http://opennlp.sourceforge.net/projects.html