

# جامعة دمشق

كلية الهندسة المعلوماتية

السنة الخامسة

قسم الذكاء الصناعي

## كشف الانتحال باستخدام الويب

الدكتور المشرف:

د. باسل الخطيب

تقديم الطلاب:

محمد صالح

مرهف فارس

محمد موسى حمد



## ملخص:

ازداد الاهتمام بأنظمة كشف الانتحال بشكل كبير خلال العقود الثلاثة الماضية. وتم اقتراح عدد من الطرق والخوارزميات لكشف الانتحال، إلا أن مشاكل عديدة لازالت تواجه العاملين في هذا المجال. من أهم المشاكل التي تواجه منهجيات كشف الانتحال في نصوص اللغات الطبيعية هو القيد بين السرعة والوثوقية أو الدقة. إضافة لذلك تقترض الأبحاث المنشورة أن فضاء البحث والمقارنة صغير نسبياً وبالتالي فإن أي خوارزمية بحث، مهما كان تعقيدها، ستعمل بشكل جيد، إلا إن هذا الافتراض يتهاوى في حالة النظم المفتوحة التي تستخدم الانترنت كمجال لبحثها. لذلك كان من الواجب البحث عن طريقة لتصغير مجال البحث قبل البدء بالمقارنة. يقدم هذا المشروع طريقة جديدة لتجاوز قيد السرعة-الوثوقية عن طريق استخدام قاعدة معطيات محلية قبل البحث في الانترنت ودمج المقارنة الدلالية مع الإحصائية. كما يقدم اقتراحاً لتصغير مجال البحث مما يجعل استخدام الانترنت كفضاء للبحث سريعاً وعملياً إلى حد ما.

## Contents

6.....	مقدمة:	1.1
7.....	تعريف الانتحال - Plagiarism:	1.2
7.....	أشكال الانتحال:	1.2.1
8.....	أمثلة عن حالات انتحال:	1.2.2
8.....	أهداف المشروع:	1.3
8.....	مجال المشروع:	1.4
9.....	لماذا المشروع؟	1.5
11.....	الأعمال ذات الصلة:	2.1
12.....	خوارزميات كشف الانتحال	2.2
12.....	كشف التشابه عن طريق تحليل أسلوب الكتابة:	2.2.1
13.....	مقارنة المستندات:	2.2.2
13.....	كشف الانتحال بالاعتماد على المعنى:	2.2.2.1
13.....	كشف الانتحال الحرفي:	2.2.2.1.1
15.....	توصيف النظام:	4.1
17.....	مراحل عمل النظام:	4.2
17.....	دخل النظام – Input:	4.2.1
17.....	التقطيع – Segmentation:	4.2.2
17.....	الفقرات الافتراضية – Virtual Paragraphing:	4.2.3
17.....	ترتيب الجمل – Segment Ranking:	4.2.4
17.....	البحث – Searching:	4.2.5
18.....	المقارنة – Comparing:	4.2.6
18.....	الإظهار – Visualization:	4.2.7
20.....	محلل النصوص Text Analyzer	.5.1
20.....	قارئ الملفات File Reader:	.5.1.1

20	المحلل النصي Parser	5.1.2
21	المجدّع Stemmer	5.1.3
21	مرتب الجمل Ranker	5.2
22	عدد الكلمات (Words Count)	5.2.1
22	أهمية الكلمات (Words Weights)	5.2.2
22	تكرار الكلمة (Word Frequency)	5.2.2.1
23	المسافة الدلالية للكلمة عن موضوع النص (Semantic Distance)	5.2.2.2
23	طول المسار (Path Length)	5.2.2.2.1
24	طريقة Wu & Palmer	5.2.2.2.2
24	الباحث Searcher	5.3
24	قاعدة معطيات محلية (Local Database)	5.3.1
25	الانترنت (The Internet)	5.3.2
25	Yahoo!	5.3.2.1
25	Google	5.3.2.2
26	المقارن Comparer	5.4
26	خوارزميات المقارنة الإحصائية (Fingerprinting Algorithms)	5.4.1
26	خوارزمية السلسلة المشتركة الأطول (LCS: Longest Common Subsequence)	5.4.1.1
27	خوارزمية Winnowing	5.4.1.2
28	نموذج فضاء الشعاعي (Vector Space Model)	5.4.1.3
29	خوارزميات المقارنة اعتماداً على المعنى (Semantic-Based Algorithms)	5.4.2
30	التشابه الدلالي (Semantic Relatedness)	5.1.1.1

# الفصل الأول:

## مقدمة

1

# 1. مقدمة:

## 1.1 مقدمة:

قسم المؤرخون والفلاسفة حياة الإنسان على كوكب الأرض إلى عدة عصور وأطلقوا على كل منها اسماً خاصاً. سمي عصرنا الحالي عصر المعلومات حيث أصبح اكتساب المعارف والحصول على المعلومات أمراً يسيراً جداً، وذلك بفضل الشبكة العنكبوتية. ولكن مع كل اختراع جديد تأتي سلبيات ومضار 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].

بدأ البحث في مجال كشف الانتحال في سبعينيات القرن الماضي، حيث بدأت الدراسات والأعمال الأولى لكشف التشابه في البرمجة وبالتحديد في جامعات علوم الحاسوب. فعلى مدى الثلاثين عاماً الماضية، تم اقتراح عدد كبير من الطرق والخوارزميات لتحديد التشابه "غير العادي" بين وظائف الطلاب.

وفي الأونة الأخيرة، اتجهت جهود الباحثين 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 والأنظمة التي تقدم خدمة الكشف عن الانتحال عبر الإنترنت.

يشرح هذا التقرير نظاماً لكشف الانتحال في نصوص اللغات الطبيعية، No More Plagiarism (NMPlagiarism). تم تطوير النظام بحيث يكون مستقل عن اللغة (Language-independent) مما يجعل إمكانية توسعته ليعالج لغات جديدة أمراً بغاية السهولة.

تتألف الخوارزمية المطورة في هذا النظام من أربع مراحل رئيسية:

1. مرحلة المعالجة الأولية Pre-processing stage: تتضمن هذه المرحلة عمليات: tokenization, stemming و stop words removal
2. مرحلة الاسترجاع Retrieval stage: يتم في هذه المرحلة استرجاع قائمة من المستندات المرشحة أن تكون مصدراً للانتحال.
3. مرحلة المقارنة Comparison stage: في هذه المرحلة تتم مقارنة المستند المشبوه مع المستندات المسترجعة وتحديد درجة التشابه بين هذه الملفات.
4. مرحلة الإظهار Visualization stage: في هذه المرحلة يتم تولين الجمل المتشابهة.

## 1.2 تعريف الانتحال - Plagiarism:

في اللغة العربية:

- "النَّحْلَةُ: الدَّعْوَى. وَأَنْتَحَلَ فَلَانٌ شِعْرَ فَلَانٍ. وَتَنَحَّلَهُ: ادَّعَاهُ وَهُوَ لغيره." [ لسان العرب]
- "وَأَنْتَحَلَهُ وَتَنَحَّلَهُ: ادَّعَاهُ لِنَفْسِهِ وَهُوَ لغيره." [ القاموس المحيط]

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

*"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."*

في عام 2001 عرّف Stuart Hannabuss 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)

### 1.2.1 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 أو المفردات بحيث يبقى المضمون نفسه و يبقى قابل للتمييز.
- 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](http://www.essaymill.com).

## 1.2.2 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

[8]"(Joseph E. Stiglitz (2006), Making Globalization Work. London: Penguin Books, p. 280.

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]

## 1.3 أهداف المشروع:

فيما يلي نبين الأهداف الرئيسية للمشروع:

1. تطوير نظام Generic لكشف الانتحال في نصوص اللغات الطبيعية حيث يكون مستقل عن اللغة Language-independent
2. البحث عن الخوارزمية والبنية الأفضل للنظام بحيث يتم تجاوز قيد السرعة-الدقة Speed-reliability.
3. البحث عن طرق جديدة لاسترجاع المستندات المرشحة لتكون مصدر الانتحال.

## 1.4 مجال المشروع:

1. يعالج NMPlagiarism نصوص اللغات الطبيعية باللغتين الانكليزية والعربية. إلا أنه، كما ذكر سابقاً، يمكن تمديده بسهولة ليشمل لغات أخرى.
2. يقوم NMPlagiarism بكشف التشابه الحرفي بين النصوص وكما يمكن كشف التشابه في حال تبديل كلمة ما بإحدى مرادفاتها.
3. تنحصر المعالجة الدلالية على المفاهيم والكلمات الموجودة في WordNet العربية والانكليزية.



## 1.5 لماذا المشروع؟

إن مشكلة الانتحال بين النصوص ليست بالجديدة حيث تم طرح العديد من الطرق للحد منها واكتشافها. ونتيجة لذلك نجد اليوم عشرات الأنظمة لكشف الانتحال بعضها مجاني والبعض الآخر تجاري، إلا أن الأنظمة المجانية لا تدعم كشف الانتحال في النصوص التي تم تعديل بعض كلماتها بمصادقاتها والأنظمة التجارية تعتبر غالية نسبياً. إضافةً لذلك، فإن دعم اللغة العربية في هذه الأنظمة لازال محدوداً بشكل أو بآخر. لذلك نقدم نظاماً مجانياً لكشف الانتحال في اللغة الانكليزية والعربية.

الفصل الثاني:

# الدراسة المرجعية

2

## 2. الدراسة المرجعية:

يعرض هذا الفصل بعض أنظمة كشف الانتحال الحالية، ومن ثم يشرح طرائق وخوارزميات كشف الانتحال المطروحة. تنتمي هذه الطرائق والخوارزميات لمجالات متعددة منها استرجاع المعطيات Information Retrieval، معالجة اللغات الطبيعية Natural Language Processing و التنقيب عن المعطيات Data Mining. إن هذا التنوع في الطرائق سببه تنوع أشكال الانتحال في نصوص اللغات الطبيعية.

### 2.1 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

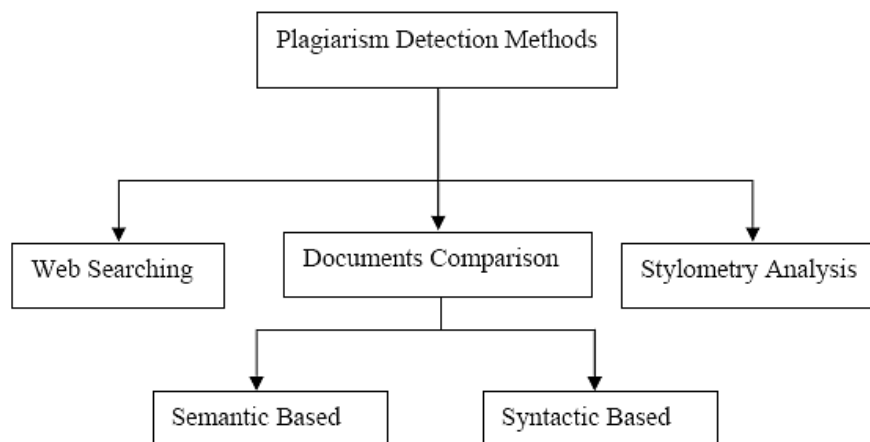
خوارزمية المقارنة	[3] Approximate string matching
نوع النظام	Java application
اللغات المدعومة	Natural languages
التكلفة	في السنة \$3,000

### أنظمة أخرى

مجانية: ChimpSky ، eTBLAST ، SeeSources ، Plagium ، CopyTracker  
تجارية: Plagiarism-detector ، Ephorus ، Copyscape ، Plagiarismdetect

## 2.2 خوارزميات كشف الانتحال

انطلاقاً من عمل Maurer وآخرين [22] فإن خوارزميات كشف الانتحال يمكن أن تصنف عموماً في ثلاثة صفوف رئيسية: الأول يحاول التعرف على أسلوب المؤلف في الكتابة وإيجاد الجمل غير المتسقة مع أسلوبه، تعرف هذه الطريقة باسم Stylometry. الصف الثاني هو مقارنة مجموعة من المستندات فيما بينها وذلك يتم تحديد الأجزاء المتشابهة. الصف الثالث يكون دخله مستند ما وتبحث عن مستندات تشابهه على شبكة الانترنت.



### 2.2.1 كشف التشابه عن طريق تحليل أسلوب الكتابة:

لا يمكننا افتراض توافر النصوص الأصلية دائماً، فربما يتم النسخ من كتاب لا يتواجد بشكل رقمي أو ربما قد يلجأ البعض لزملائهم في كتابة الوظائف. في مثل هذه الحالات نلجأ إلى كشف الانتحال عن طريق تحليل أسلوب الكتابة حيث أن الخوارزميات الأخرى تفشل في كشفه. في هذه الطريقة نقوم بتطبيق خوارزمية كشف الانتحال على أكثر من عمل (مقالة، وظيفة، كتاب) لنفس الكاتب من دون اللجوء إلى مصادر خارجية وتسمى هذه الطريقة Intrinsic plagiarism detection methods [22,23].

إن تحليل أسلوب الكتابة هو طريقة إحصائية للتعرف على مؤلف نص ما، ولكن هذه الطريقة تتطلب بعض السمات اللغوية المعرفة بدقة عالية. تعتمد هذه الطريقة على الفرضية القائلة أن لكل مؤلف أسلوب كتابة خاص به فإذا تغير هذا الأسلوب بين جملة وأخرى أو بين فقرة وأخرى فذلك يعني أن النص قد يكون منتهكاً [24]. فمثلاً أحد أشكال تغير الأسلوب هو التبديل في استخدام الضمائر "We/our" و "I/my" أو عندما يتم استخدام أحرف الجر وأدوات التعريف بطرق مختلفة بشكل كبير.

تصنف طرق تحليل أسلوب الكتابة حسب حجم الوحدة المستخدمة في تحديد الأسلوب، وبالتالي يمكن وضع هذه الطرق ضمن خمسة صفوف [23]: (1) التحليل على مستوى المحرف (طريقة احصائية) (2) التحليل على مستوى الجملة (3) التحليل على مستوى أجزاء الكلام (4 Part-of-speech) التحليل عن طريق متابعة مجموعة محددة من الكلمات Closed-class word (5) التحليل على مستوى هيكلية النص كاملاً.

إن هذه الطريقة لم تنتشر كثيراً وذلك لصعوبة إثبات أن نصاً ما منتحلاً من دون إظهار النص الأصلي كدليل. ولكن يمكن استخدام هذه الطريقة كمؤشر على المستندات التي يمكن أن تكون منتحلة وبالتالي نخضعها لمزيد من المقارنة والبحث [22,23].

## 2.2.2 مقارنة المستندات:

إن الهدف الرئيسي لنظام كشف الانتحال هو توضيح الانتهاكات في حقوق الملكية. وكما ذكرنا سابقاً، فإن انتهاكات حقوق الملكية لها أشكال عديدة منها النسخ الحرفي للجمال والفقرات وهذا يمكن اكتشافه باستخدام الخوارزميات الإحصائية. ولكن يجب ألا ننسى أن في اللغات الطبيعية تعقيداً كبيراً، حيث نجد لكل مفهوم أكثر من كلمة تعبر عنه، أي باختصار يوجد لكل الكلمات مرادفات وهذا يسهل عملية الانتحال ويصعب اكتشافها! ولذلك نحن بحاجة لخوارزميات معالجة اللغات الطبيعية لنكشف الانتحال في النصوص التي تم تبديل كلماتها بمرادفاتها.

### 2.2.2.1 كشف الانتحال بالاعتماد على المعنى:

إن معظم أنظمة كشف الانتحال تعتمد على المقارنة الحرفية للنصوص ولا تهتم بمعنى بالنصوص، وكنيجة لذلك فإن بعض التحوير في النص يؤدي إلى فشل النظام. إن التعديل في النص عن طريق تبديل الكلمات بمرادفات يمكن معالجته باستخدام قاموس مرادفات مثل WordNet، إلا أن المسألة ليست بهذه البساطة فاللغات الطبيعية تمتاز بالغموض مما يجعل اختيار المعنى المناسب لكلمة ما ليس بالأمر السهل [3]. سنأتي على شرح هذه الطريقة بالتفصيل لاحقاً.

#### 2.2.2.1.1 كشف الانتحال الحرفي:

على عكس الطريقة السابقة، فإن هذه الطريقة لا تهتم بمعنى الكلمة أبداً، فمثلاً كلمتي "الجشع" و "البخل" تعتبران مختلفتين. يعبر المثال السابق، وبلا شك، عن محدودية هذه الطريقة إلا أنها في الوقت نفسه تسرع عملية المقارنة بشكل كبير على عكس الطريقة السابقة. سنتطرق إلى تفاصيل بعض الطرائق في هذا الصف لاحقاً.

الفصل الثالث:

# البنية العامة للنظام

3

### 3. البنية العامة للنظام

في هذا الفصل نشرح العوامل الرئيسية التي تحدد أي نظام لكشف الانتحال، وكيف تم تطبيقها في نظام NMPlagiarism. ومن ثم نبين مراحل عمل النظام.

#### 3.1 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):

العامل	الدالة	NMPlagiarism
مجال البحث	مجال البحث يمكن أن يكون: الانترنت، باستخدام محركات البحث، قواعد معطيات خاصة محلية. يقابل هذا التقسيم النوعين التاليين: Open-system: البحث في الانترنت باستخدام محركات البحث Local-system: البحث في قواعد معطيات محلية.	يمكن وصف NMPlagiarism بأنه hyper-system لأنه يستخدم قاعدة معطيات محلية والانترنت لكشف عملية الانتحال في نصوص اللغات الطبيعية.
زمن التحليل	الزمن بين لحظة تقديم المستند للفحص و لحظة ظهور النتائج. أي الزمن اللازم لتحديد فيما إذا كان نص منتحل أم أصيل.	يعتمد زمن التحليل على الخوارزمية المستخدمة في المقارنة، فمثلاً بالنسبة للخوارزميات الإحصائية فإن الزمن اللازم هو: 55~ ثانية لنص مكون من 2000 حرف.
عمق الاختبار	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	خوارزمية المقارنة المستخدمة، خوارزميات إحصائية، خوارزميات تعتمد على المعنى....	الإحصائية Fingerprint-based system و Winnowing و Vector LCS و space model الدلالية: Semantic relatedness
الدقة	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.	High precision, 94% High recall, 97%

#### Local-System و 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. بالنسبة لـ NMPlagiarism فإن البحث في قاعدة معطيات محلية يحسن من أداء النظام بشكل ملموس، والبحث في الإنترنت، إن لم يجد النتيجة في قاعدة المعطيات، يسحن من دقته. لذلك فإن NMPlagiarism تجاوز مشاكل Open-system وال Local-system بالدمج بينهما.

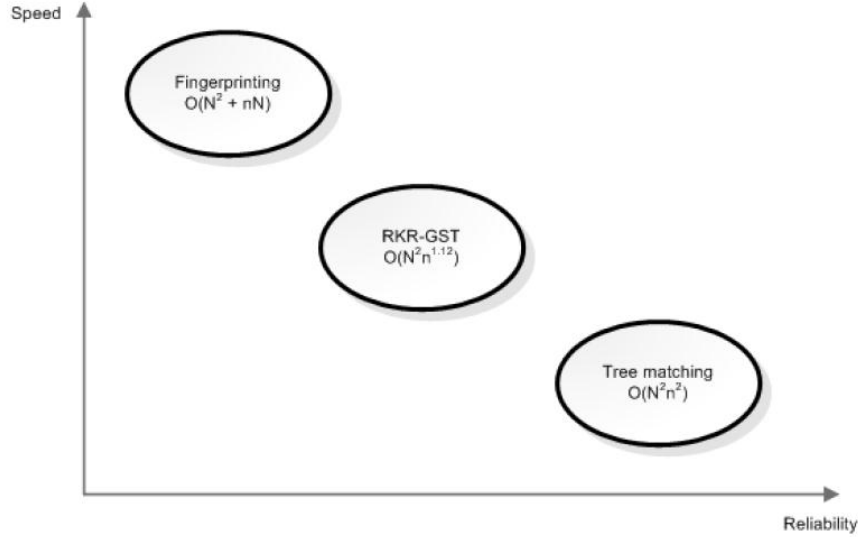
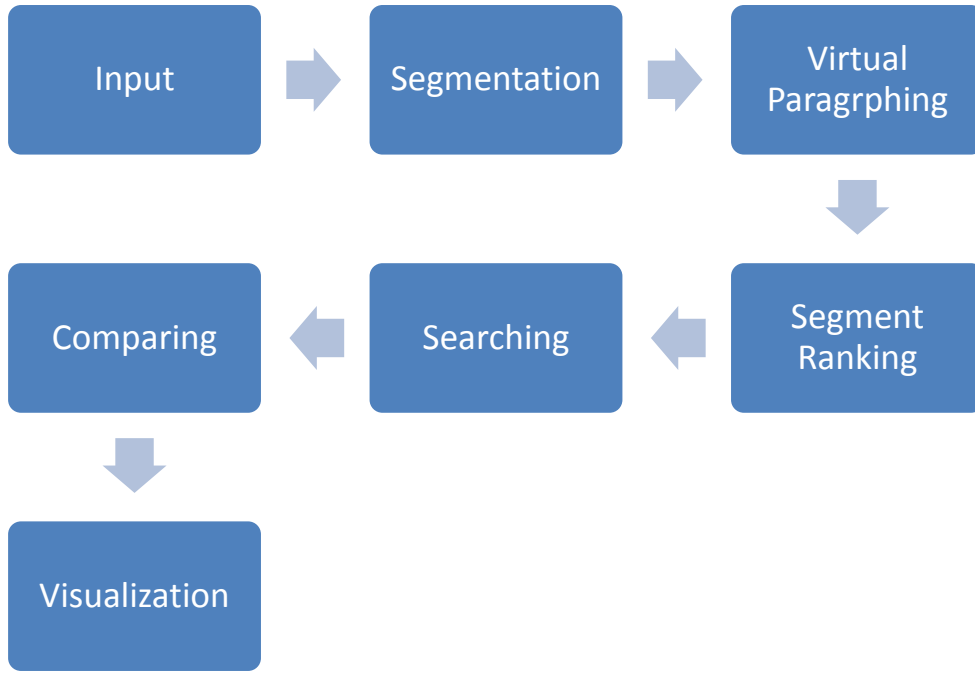


Fig. 6.1. Speed and reliability of different plagiarism detection schemes [4]



### 3.2. مراحل عمل النظام:



#### 3.2.1 دخل النظام – Input:

دخل النظام هو عبارة عن سلسلة نصية أو ملف مهيكّل بأحد الصيغ التالية: MS Word, PDF, HTML, MS PowerPoint

#### 3.2.2 التقطيع – Segmentation:

في هذه المرحلة يتم تقطيع النص إلى سلسلة من الجمل كل جملة منها مؤلفة من سلسلة من الكلمات.

#### 3.2.3 الفقرات الافتراضية – Virtual Paragraphing :

في هذه المرحلة يتم تشكيل فقرات افتراضية مؤلفة من عدد ثابت من الجمل، حيث يتم تحديد عدد الجمل في الفقرة من قبل المستخدم. تم إطلاق صفة افتراضية على هذه الفقرات لأنها تختلف عن فقرات المستند الأصلية.

#### 3.2.4 ترتيب الجمل – Segment Ranking :

في هذه المرحلة يتم توزيع جمل الفقرات الافتراضية حسب أوزان كلمات هذه الجمل. حيث يعبر وزن الجملة عن مدى ارتباطها وتعبيرها عن الفقرة الافتراضية التي تنتمي إليها. وبعد ذلك يتم ترتيب الجمل حسب أوزانهم تنازلياً.

#### 3.2.5 البحث – Searching :

في مرحلة البحث يستخدم النظام الجمل المرتبة من المرحلة السابقة للبحث عن نصوص مشابهة للنص المدخل ومرشحة أن تكون مصدر الانتحال. يتم البحث أولاً في قاعدة المعطيات المحلية حيث تمثل كل جملة استعلاماً واحداً

وعند الاستعلام عن جميع الجمل وعدم الوصول إلى نتيجة يقوم النظام بالبحث في الانترنت باستخدام أحد محركات البحث.

### 3.2.6 المقارنة – Comparing :

في هذه المرحلة يقارن النظام المستند النصي المدخل مع الملفات التي استرجعها في الخطوة السابقة، سواء من الانترنت أو قاعدة المعطيات المحليّة. تتم المقارنة باستخدام الخوارزميات الإحصائية أو الدلالية أو كليهما معاً.

### 3.2.7 الإظهار – Visualization:

أخيراً، في هذه المرحلة يظهر النظام النتائج عن طريق تلوين الأجزاء المتشابهة بين النص المدخل والنصوص المسترجعة.

الفصل الرابع:

# بنية NMPlagiarism

4

## 4. بنية النظام

يتألف NMPlagiarism من أربع وحدات رئيسية خصصت لكل منها مهمة، وقد تم بناء هذه الوحدات بعناية بحيث يمكن توسيع أي منها بإضافة توابع جديدة أو تعديلها باستبدال توابع قديمة دون التأثير في بعضها البعض.



### 4.1. محلل النصوص Text Analyzer



تقوم هذه الوحدة بقراءة الملفات من الأنواع المدعومة في NMPlagiarism وهيكله هذه الملفات المقروءة وفق هياكل محددة بهدف تبسيط معالجتها لاحقاً في باقي وحدات النظام. وتتألف هذه الوحدة من ثلاث كتل أساسية:

#### 4.1.1. قارئ الملفات File Reader:

تستخدم كتلة قراءة الملفات بعض المكتبات لقراءة المحتوى النصي لعدة أنواع من الملفات، ويستخدم ذلك المحتوى كدخل للكتلة التالية. يدعم NMPlagiarism أنواع الملفات النصية الأكثر شهرة واستخداماً: Text, MS Word, PDF, HTML.

وانطلاقاً من البنية المرنة للنظام يمكن زيادة أنواع الملفات المدعومة ببساطة وذلك بإضافة مكتبة جديدة لهذه الوحدة تدعم النوع الجديد من الملفات النصية المراد إضافتها لمجموعة الملفات المدعومة في NMPlagiarism.

#### 4.1.2. المحلل النصي Parser:

تعالج هذه الكتلة المحتوى النصي للملفات - خرج الكتلة السابقة - والذي هو نص غير مهيكّل Plain Text لتشكّل منه مجموعة من المقاطع النصية والتي تعتبر خرج هذه الكتلة. يتكوّن كل مقطع من هذه المقاطع من قائمة من الجمل والتي بدورها تحوي قائمة من الكلمات.

تستخدم هذه الكتلة المكتبة OpenNLP (مركز تنظيمي للمشاريع المفتوحة المصدر المتعلقة بمعالجة اللغات الطبيعية) والتي بدورها تستخدم الـ WordNet Semantic Net لتقسيم الدخل النصي إلى جمل، ولتجزئة ذلك الدخل إلى قائمة من الكلمات (Tokens). يوجد حالياً العديد من الأدوات و المشاريع تحت اسم OpenNLP استخدمنا منها OpenNLP Tokenizer و OpenNLP Splitter.

### 4.1.3. المجدع Stemmer:

التجذيع هو عملية إزالة جزء من السوابق واللواحق وبعض الحروف غير الأصلية في الكلمة بهدف إيجاد جذع هذه الكلمة. ولا يجب أن يطابق جذع كلمة ما الجذع المعجمي لهذه الكلمة، وإنما يكفي أن يكون لمجموعة من الكلمات المترابطة نفس الجذع وإن لم يكن هذا الجذع كلمة صحيحة، وذلك بهدف معالجة مجموعة الكلمات تلك ككلمة واحدة ممثلة بالجذع مما يساهم تحسين النتائج وتقليل عدد عمليات المعالجة.

مثال: جذع الكلمات "يذهبون" و"ذاهبون" هو "ذهب".

ويختلف التجذيع عن التجذير وهو عملية إعادة الكلمة إلى جذرها في المعجم، وقد يضيع في هذه العملية جزء كبير من معنى الكلمة.

مثال: جذر الكلمات "منافسة" و"نفيضة" هو "نفس" رغم اختلاف معناها.

### 4.2. مرتب الجمل Ranker



للبحث عن الملفات المشابهة لملف الدخل كمصادر لهذا الملف يستخدم NMPlagiarism قاعدة معطيات محلية ويستعين ببعض محركات البحث المعروفة والموثوقة. وكل من قاعدة المعطيات أو محركات البحث يحتاج لاستعلام (Query) للبحث عنه، ولما كان من الممكن للملف الدخل أن يكون كبير الحجم فإن استخدام النص الكامل المحتوى في هذا الملف للبحث عملية مكلفة جداً ولا تضمن نتائج جيدة خاصة في حال تعديل هذا الملف عن مصادره. يختار NMPlagiarism تلقائياً مجموعة جزئية من الجمل لاستخدامها في البحث عن المصادر، وهذا مما يميز هذا النظام عن غيره من أنظمة كشف الانتحال في النصوص، حيث أن معظم الأنظمة الحالية تطلب من المستخدم تحديد جملة البحث أو موضوع النص (رياضيات، علوم، تاريخ..).

يقوم مرتب الجمل بترتيب جمل المحتوى النصي للملف الدخل تبعاً لأوزان هذه الجمل. يعبر وزن جملة ما عن أهمية هذه الجملة ضمن المحتوى النصي ومقدار المعلومات التي من الممكن أن تحتويه عن هذا المحتوى. ولحساب هذه الأوزان يمكن لـ NMPlagiarism استخدام إحدى التوابع التالية:

1. تحديد أهمية الجملة حسب عدد الكلمات فيها [2].
2. تحديد أهمية كلمات النص و من ثم تحديد أهمية الجملة حسب أهمية كلماتها.
- تختلف استراتيجيات تحديد أهمية الكلمة في النص منها:
- (1) تحديد أهمية الكلمة حسب عدد مرات ورودها في النص فالكلمة الأقل وروداً تحمل قيمة أكثر [6]
- (2) تحديد أهمية الكلمة بتابع يعتمد على تابع التوزع الطبيعي
- (3) المسافة الدلالية للكلمة عن موضوع النص.

#### 4.2.1. عدد الكلمات (Words Count):

وفيه يتم تحديد وزن الجملة حسب عدد كلماتها، ويمكن لعدد كلمات جملة ما أن يكون العدد الفعلي لجميع كلمات الجملة أو عدد الكلمات غير المصنفة من الكلمات كثيرة الاستخدام في اللغة (Stop Words) مثل أحرف الجر وأحرف العطف وغيرها، وذلك لأن كلمات كهذه قد لا تضيف أية معلومات لجملة ما.

ويمكن التعبير عن هذا التابع بالعلاقة الرياضية:

$$w(s) = a \cdot x$$

حيث: يمثل المتحول  $s$  الجملة المراد احتساب وزنها، و  $x$  يمثل عدد كلمات هذه الجملة (مع أو بدون احتساب الكلمات غير المصنفة من الكلمات كثيرة الاستخدام)، و  $a$  يمثل ثابت أكبر تماماً من الصفر.

وتتصف توابع التوزين بدلالة عدد الكلمات ببساطة التحقيق وسرعة التنفيذ، إلا أن نتائج هذه التوابع غير جيدة مقارنة بتوابع التوزين بدلالة أهمية الكلمات.

#### 4.2.2. أهمية الكلمات (Words Weights):

وفيه يتم تحديد وزن الجملة بدلالة أوزان كلماتها، حيث يتم احتساب وزن لكل كلمة من كلمات الجملة يعبر أيضاً عن أهمية هذه الكلمة ضمن المحتوى النصي وكمية المعلومات التي من الممكن أن تحتويها هذه الكلمة عن ذلك المحتوى. وكما في توابع عدد الكلمات يمكن احتساب أهمية جميع كلمات الجملة أو أهمية الكلمات غير المصنفة من الكلمات كثيرة الاستخدام.

ويمكن التعبير عن هذه التوابع بالعلاقة الرياضية:

$$w(s) = f(v(1), v(2), \dots, v(i))$$

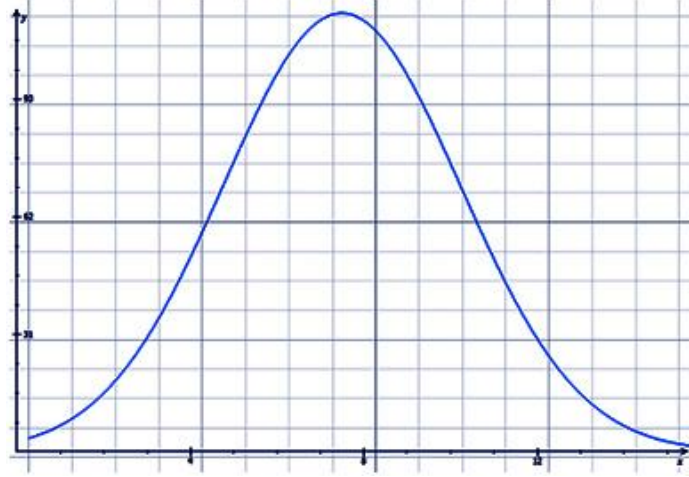
حيث: يمثل المتحول  $s$  الجملة المراد احتساب وزنها، و  $f$  تابع لاحتساب وزن الجملة بدلالة أوزان كلماتها ومن أشهر هذه التوابع تابع المتوسط الحسابي (mean)، و  $v(i)$  يمثل وزن الكلمة  $i$ .

وتعتمد هذه التوابع في جودتها وزمن تنفيذها بشكل أساسي على تابع توزين الكلمات، إلا أنها وبشكل عام تتصف بجودة النتائج وزمن التنفيذ الكبير مقارنة بتوابع التوزين بدلالة عدد الكلمات.

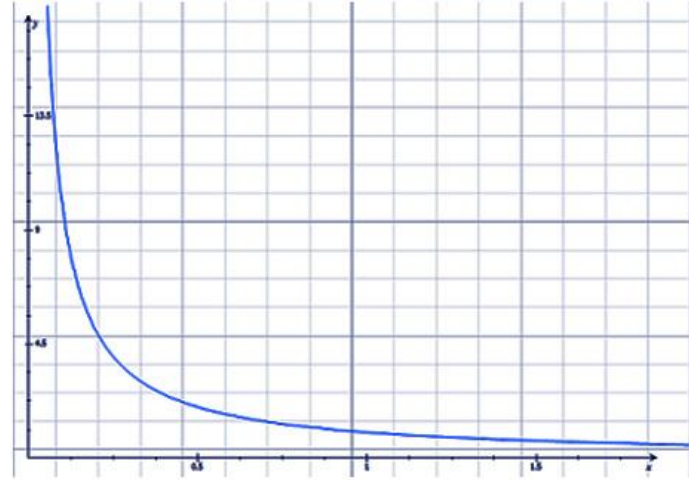
ولاحتساب وزن كلمة ما ضمن الجملة أو المحتوى النصي بشكل عام يمكن تطبيق إحدى التوابع التالية:

##### 4.2.2.1. تكرار الكلمة (Word Frequency):

حيث تعطى الكلمة الوزن بحسب عدد مرات ظهورها ضمن المحتوى النصي. وبحسب نظرية المعلومات فإن الكلمات ذات التكرار الأقل و ذات التكرار الأكبر هي الكلمات الأقل أهمية، في حين تكون الكلمات ذات التكرار الأقرب ما يمكن للقمة في تابع التوزع الطبيعي لتكرار الكلمات هي الأكثر أهمية.



كما يمكن في بعض الأحيان اعتبار الكلمات الأقل تكراراً هي الكلمات المهمة، إلا أن هذا الافتراض لا يعطي نتائج بجودة الافتراض السابق.



#### 4.2.2.2. المسافة الدلالية للكلمة عن موضوع النص (Semantic Distance):

وفيها يجب تحديد موضوع المحتوى النصي سواء من قبل المستخدم أو باستخدام نظام تصنيف خارجي (Text Classification System). وبناء عليه يتم تحديد أهمية الكلمة بحسب تشابهها الدلالي مع موضوع المحتوى النصي وذلك باستخدام شبكة دلالية تمثل العلاقات بين معظم كلمات اللغة. ومن أهم هذه الشبكات وأكثرها استخداماً هي الـ WordNet Ontology. وتعتبر الكلمات الأكثر تشابهاً مع الموضوع هي الأكثر أهمية وبالتالي ذات الوزن الأكبر. وهناك عدة طرق لحساب التشابه بين كلمتين في شبكة دلالية نذكر من أهمها:

##### 4.2.2.2.1. طول المسار (Path Length):

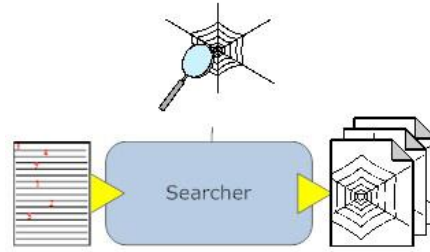
وفيها يكون التشابه مقلوب عدد العقد على طول المسار الأقصر بين العقدتين الممثلتين للكلمتين المراد احتساب تشابههما.

#### 4.2.2.2. طريقة Wu & Palmer:

تقيس هذه الطريقة التشابه بحساب عمق كل من العقدة الممثلة للكلمة الأولى والعقدة الممثلة للكلمة الثانية والعقدة الممثلة لأول أب مشترك (LCS: Least Common Subsumer) بين تلك العقدتين في الشبكة الدلالية، وبحسب التشابه بالعلاقة:

$$\text{Similarity}(w1, w2) = 2 * \text{depth}(\text{LCS}) / (\text{depth}(w1) + \text{depth}(w2))$$

### 4.3. الباحث Searcher



تستخدم هذه الوحدة أهم الجمل الناتجة عن مرتب الجمل للبحث عن الملفات النصية المشابهة للمحتوى النصي للملف الدخل لاستخدامها والملف الدخل كدخل لوحدة المقارنة. يمكن للباحث استخدام قاعدة معطيات محلية أو الانترنت كفضاء البحث.

#### 4.3.1. قاعدة معطيات محلية (Local Database):

لاستخدام قاعدة معطيات محلية كمجال البحث يجب أن تضم قاعدة المعطيات هذه كمية كبيرة جداً من الملفات النصية كما يجب أن تضم آخر المقالات المنشورة للمحافظة على جودة النتائج. وعليه يقوم NMPiagiarism بتخزين جميع الملفات النصية التي يقوم بالمقارنة معها عند استخدام الانترنت كمجال للبحث لتستخدم هذه الملفات فيما بعد في حال الرغبة في استخدام قاعدة المعطيات المحلية كمجال للبحث. وبالرغم من المجازفة في جودة النتائج في حال استخدام قاعدة معطيات محلية وذلك في حال كانت الملفات المصدر لم يتم تخزينها إلى الآن إلا أن استخدامها يوفر سرعة كبيرة جداً في التنفيذ مقارنة باستخدام الانترنت للبحث.

ويتم البحث في قاعدة معطيات محلية باستخدام تقنية البحث ضمن نظام إدارة قواعد المعطيات Microsoft SQL Server والمعروفة بـ Full-Text Search وذلك باستخدام أية جملة يراد البحث عنها (Search Query).

تتيح Full-Text Search أرشفة (Indexing) سريعة ومرنة للاستعلام باستخدام كلمات مفتاحية عن نصوص مخزنة في قاعدة معطيات ضمن نظام إدارة قواعد المعطيات Microsoft SQL Server. تنفذ الاستعلامات باستخدام Full-Text Search عمليات بحث لغوية في البيانات المخزنة وذلك بالعمل على الكلمات والعبارات استناداً إلى قواعد لغة معينة.

يمكن لـ Full-Text Search توليد قيمة اختيارية (أو وزن) تعبر عن مدى ارتباط البيانات الممثلة لنتائج البحث بالاستعلام المستخدم كجملة البحث. وبحسب هذا الوزن من أجل كل نص في قاعدة المعطيات، ويمكن استخدامه كمعيار لترتيب لفرز مجموعة نتائج الاستعلام حسب ارتباطها بهذا الاستعلام.

وتستند عملية احتساب الأوزان على علاقة التصنيف المعروفة بـ OKAPI BM25، والموضحة بشكل تفصيلي في Appendix A.



### 4.3.2. الانترنت (The Internet):

لاستخدام الانترنت كمجال للبحث لابد من الاستعانة بمحركات البحث في الانترنت، يمكن في NMPlagiarism الاستعانة بأحد محركات البحث التالية:

#### 4.3.2.1. Yahoo!:

يوفر محرك البحث YAHOO! العديد من واجهات برمجة التطبيقات (API) لعدة استخدامات، ولعل من أهمها واجهة البحث (Yahoo Search API) التي تتيح للمطورين استخدام وظائف محرك البحث في Yahoo!. وللاستخدام الأمثل لهذه الواجهة يجب الحصول على مفتاح لها (Application Key) والذي يمكن طلبه من YAHOO!، كما يجب العمل على تحقيق بعض الوظائف والمتطلبات من قبل المطور للحصول على النتائج المتوقعة. معظم واجهات برمجة التطبيقات (API) التي تقدمها YAHOO! تعيد نتائج مهيكلة وفق الصيغة القياسية XML مما يجعل معالجتها أبسط.

ويعتبر الرابط (URL) التالي كمثال لاستخدام واجهة البحث المقدمة من YAHOO!:

<http://api.search.yahoo.com/WebSearchService/V1/webSearch?appid=PlagiarismDetectionSystem&query=plagiarism&results=2>

حيث:

V1: تمثل رقم إصدار واجهة البحث (API Version).  
webSearch: تمثل نوع واجهة برمجة التطبيق (API) المستخدمة.  
appid: يمثل مفتاح التطبيق الخاص بالواجهة المستخدمة والذي يمكن الحصول عليه بطلبه من YAHOO!.  
query: تمثل النص المراد البحث عنه في الانترنت.  
results: تمثل عدد النتائج المرادة من محرك البحث.  
ويعيد هذا الرابط ملف XML يحوي نتائج البحث ممثلة برابط وجزء من محتويات هذا الرابط من أجل كل نتيجة من النتائج.

#### 4.3.2.2. Google:

تقدم Google أيضاً العديد من واجهات برمجة التطبيقات (API) ومنها واجهة البحث. ويمكن لهذه الواجهة أن تكون JavaScript-Based (حيث تتم إضافة رماز مصدري إلى صفحة HTML معينة لاستخدام وظائف محرك البحث في Google)، كما يمكن أن تكون REST-Based كما كانت واجهة البحث المقدمة من YAHOO! (حيث يتم التخاطب مع المخدم الخاص بواجهة البحث باستخدام رابط معين، ويمكن الحصول على النتائج باستخدام التابع HTTP Get). وتهيكّل Google نتائج واجهة البحث وفق الصيغة القياسية JSON.

يتطلب استخدام واجهة البحث المقدمة من Google أيضاً مفتاح تطبيق خاص بها (Application Key) يطلب من Google، ويتم استخدامه كما تم استخدام مفتاح التطبيق في الواجهة الخاصة بـ YAHOO!.

#### 4.4. المقارن Comparer

تقوم هذه الوحدة بمقارنة المحتوى النصي لدخل النظام مع مجموعة النصوص نتائج البحث، أو مقارنة مجموعة ملفات نصية فيما بينها. وينتج عن هذه الوحدة الخرج النهائي للنظام الممثل بقائمة من التطابقات (أو التشابهات) بين النصوص دخل هذه الوحدة. ويعتمد أداء هذه الوحدة على خوارزمية المقارنة المستخدمة والتي يمكن تصنيفها بشكل عام إلى خوارزميات إحصائية (Statistical Algorithms) وخوارزميات معتمدة على المعنى أو خوارزميات دلالية (Semantic-Based Algorithms).

يبين الشكل التالي العلاقة بين جودة (وثوقية) نتائج خوارزميات المقارنة من جهة وزمن تنفيذ (تعقيد) هذه الخوارزميات من جهة أخرى. ومن السهل ملاحظة أن تعقيد الخوارزميات الإحصائية أقل بكثير من الخوارزميات المعتمدة على المعنى، في حين أن جودة نتائج الأخيرة أكبر بكثير منها في الأولى. ومن هنا تبرز صعوبة تحديد أي من الخوارزميات أفضل للتطبيق بسبب العلاقة التبادلية الدائمة بين الوثوقية من جهة وزمن التنفيذ من الجهة الأخرى.

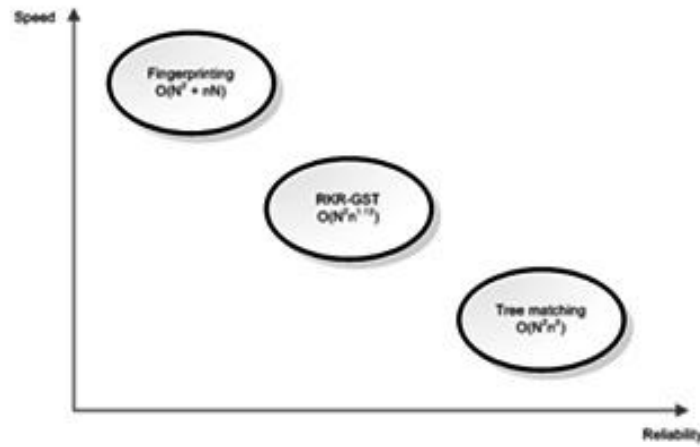


Fig. 1. Speed and reliability of different plagiarism detection schemes [4]

يمكن في NMPlagiarism استخدام إحدى خوارزميات المقارنة التالية:

##### 4.4.1. خوارزميات المقارنة الإحصائية (Fingerprinting Algorithms):

###### 4.4.1.1. خوارزمية السلسلة المشتركة الأطول (LCS: Longest Common Subsequence):

تهدف هذه الخوارزمية لإيجاد أطول سلسلة جزئية مشتركة بين مجموعة من السلاسل (عادة تتألف هذه المجموعة من سلسلتين). وتعتبر هذه الخوارزمية من الخوارزميات التقليدية في مقارنة السلاسل النصية لقياس درجة التشابه أو حتى الاختلاف بينها.

إن السلسلة الجزئية الناتجة هي سلسلة تظهر في جميع السلاسل مرتبة بنفس الترتيب النسبي لعناصر السلسلة، ولكن لا يشترط التالي المباشر لعناصر هذه السلسلة في جميع السلاسل. مثلاً: جميع السلاسل "123"، "124"، "245" هي سلاسل جزئية من السلسلة "12345".

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

#### • السلسلة المشتركة الأطول لسلسلتين:

تتصف مسألة السلسلة المشتركة الأطول بالبنية الأمثلية (Optimal Substructure)، حيث يمكن تقسيم المسألة إلى مسائل أصغر وأبسط وكل منها قابل للتقسيم لمسائل أكثر بساطة وهكذا حتى يصبح حل المسائل الجزئية واضح. كما تتصف هذه المسألة أيضاً بتداخل المسائل الجزئية (Overlapping Sub-Problems)، حيث يعتمد حل مسألة في مستوى ما على مجموعة حلول مسائلها الجزئية في المستوى الأدنى. إن المسائل ذات الخواص السابقة (Optimal Substructure and Overlapping Sub-Problems) يمكن حلها بواسطة البرمجة الديناميكية، حيث يتم بناء الحل بدءاً من حلول أبسط المسائل الجزئية. وتتطلب هذه العملية تذكر الحلول الجزئية (Memoization)، حيث يتم حفظ حلول المسائل الجزئية من مستوى معين في جدول بحيث تكون متاحة لاستخدامها في حل المسائل الجزئية في المستويات الأعلى.

#### • تعريف تابع السلسلة المشتركة الأطول:

لنعرف سلسلتين بالشكل التالي:  $X = (x_1, x_2, \dots, x_m)$  و  $Y = (y_1, y_2, \dots, y_n)$ .

ولنعرف مجموعة سوابق السلسلة  $X$  بالشكل  $X_1, X_2, \dots, X_m$ ، ومجموعة سوابق السلسلة  $Y$  بالشكل  $Y_1, Y_2, \dots, Y_n$ .

يمثل التابع  $LCS(X_i, Y_j)$  مجموعة السلاسل المشتركة الأطول بين مجموعتي السوابق  $X_i$  و  $Y_j$ ، ويحسب بالعلاقة:

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

وبالتالي لإيجاد السلسلة المشتركة الأطول لـ  $X_i$  و  $Y_j$ ، نقارن العنصرين  $x_i$  و  $y_j$ ، ففي حال كانا متساويين تتم إضافة هذا العنصر  $(x_i)$  لنهاية السلسلة  $LCS(X_{i-1}, Y_{j-1})$ ، وفي حال كانا غير متساويين تعاد السلسلة الأطول بين السلسلتين  $LCS(X_i, Y_{j-1})$  و  $LCS(X_{i-1}, Y_j)$  كخرج.

#### 4.4.1.2 خوارزمية Winnowing:

تعتبر هذه الخوارزمية من خوارزميات البصمة (Fingerprinting Algorithms) والتي تعمل على مقابلة سلسلة كبيرة من البيانات بسلسلة رقمية ذات حجم أصغر بكثير، تعرف هذه السلسلة صغيرة الحجم ببصمة سلسلة البيانات، حيث يفترض أن تمثل سلسلة البيانات الأصلية بشكل وحيد كما تعرف بصمة الإنسان صاحبها بشكل وحيد. وتستخدم خوارزميات البصمة في عمليات مقارنة بيانات ذات حجوم كبيرة أو في عمليات إرسالها عبر الشبكات الحاسوبية.

تقسم خوارزمية Winnowing سلسلة البيانات (النصوص في NMPPlagiarism) إلى سلاسل جزئية متساوية الطول وطول كل منها  $k$  وعناصر هذه السلاسل متتالية مباشرة في سلسلة البيانات الأصلية وهذا ما يعرف بـ  $k$ -grams. إن عدد هذه السلاسل الجزئية مساوٍ لعدد عناصر سلسلة البيانات الأصلية منقوصاً منه طول السلسلة الجزئية. حيث يتم اعتبار أول  $k$  عنصر كأول سلسلة جزئية ثم تتم إزاحة هذه السلسلة بمقدار عنصر واحد لتتشكل لدينا سلسلة جزئية جديدة وهكذا حتى آخر عنصر في سلسلة البيانات. ثم يتم ترميز كل مجموعة من المجموعات الجزئية برقم (hash) ويجب أن يميز هذا الرقم مجموعته الجزئية بشكل وحيد، وتكون بصمة سلسلة البيانات مجموعة جزئية من الأرقام (الرموز) الناتجة عن ترميز جميع المجموعات الجزئية.

وتختلف خوارزميات البصمة فيما بينها بشكل أساسي في اختيار المجموعة الجزئية الأخيرة المعبرة عن بصمة سلسلة البيانات، في حين تتشابه فيما بقي من مراحل العمل الأساسية كاحتساب الـ  $k$ -grams وترميز كل منها.

لتكن سلسلة الرموز  $h_1, h_2, \dots, h_n$  الممثلة لسلسلة البيانات المراد احتساب بصمتها والمشكلة بترميز كل من الـ  $k$ -grams في سلسلة البيانات تلك. ولنعرف النافذة  $w$  على سلسلة الرموز. كل موقع  $1 \leq i \leq n-w+1$  في سلسلة الرموز يعرف نافذة من تلك الرموز  $h_i \dots h_{i+w-1}$ . تختار Winnowing رمز واحد من كل نافذة من النوافذ السابقة ليكون في بصمة سلسلة البيانات الأصلية وذلك وفق الاستراتيجية التالية:

في كل نافذة تختار الخوارزمية الرمز ذو القيمة الأصغر، وفي حال كانت هناك عدة رموز تحمل أصغر قيمة يتم اختيار الرمز في أقصى اليمين. وفي حال كان الرمز ذو القيمة الأصغر النافذة الحالية وحيد و قيمته مساوية لقيمة آخر رمز تم اختياره ليكون ضمن بصمة سلسلة البيانات يتم تجاهل هذا الرمز دون إضافته للبصمة والانتقال مباشرة للنافذة التالية.

إن هذه الطريقة تعتمد على ما لوحظ تجريبياً بأنه غالباً ما تبقى القيمة الصغرى في النافذة الحالية هي نفسها في النافذة اللاحقة، وعليه ستكون سلسلة البصمة أقصر بكثير من سلسلة البيانات الأصلية.

بعد إيجاد بصمة كل نص من النصوص المراد مقارنتها يتم تطبيق إحدى خوارزميات المقارنة على سلاسل البصمات لمعرفة التشابه، كأن نطبق خوارزمية السلسلة المشتركة الأطول (LCS) والموضحة سابقاً.

مثال:

ليكن لدينا النص التالي:

A do run run run, a do run run

1- حذف الفراغات والمحارف الخاصة (White Spaces).

Adorunrunrunadorunrun

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

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

3- ترميز السلاسل الجزئية السابقة (Hashing).

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

4- تقسيم سلسلة الرموز إلى نوافذ من 4 عناصر لكل منها.

(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 بواسطة خوارزمية Winnowing:

17 17 8 39 17

#### 4.4.1.3 نموذج فضاء الشعاعي (Vector Space Model):

هو نموذج رياضي (جبري) لتمثيل المستندات النصية. حيث يتم تمثيل المحتوى النصي لكل من الملفات بشعاع من عدة مركبات، وتحسب قيم هذه المركبات بدلالة كلمات المحتوى النصي.

$$x = (x_1, x_2, \dots, x_n)$$

تمثل مركبات هذا الشعاع سمات النص (Features)، حيث يتم تعريف قائمة من الكلمات تعتبر سمات في النصوص، وترتبط كل سمة بكلمة معينة وتعبّر عنها. عادة ما تأخذ السمة في نص معين إحدى قيمتين:

- قيمة ثنائية تبين فيما إن وردت الكلمة المرتبطة بهذه السمة في النص أم لا.
- قيمة حقيقية تمثل وزن الكلمة المرتبطة بهذه السمة في النص، ويقاس هذا الوزن بإحدى خوارزميات توزيع الكلمات الموضحة سابقاً، وعادة ما يستخدم تكرار الكلمة (Term Frequency) ضمن النص كقيمة السمة المرتبطة بتلك الكلمة في نموذج الفضاء الشعاعي.

مثال: ليكن لدينا النصين التاليين:

$x1$ : "All-star game will be held in Boston"

$x2$ : "Chess is the champion of games"

يمثل النصين السابقين وباستخدام القيم الثنائية للسّمات وباعتبار مجموعة السّمات تتألف من الكلمات الأربع التالية (-all- "star", "Boston", "chess", "game" بالشعاعين التاليين:

$$x1 = (1, 1, 0, 1)$$

$$x2 = (0, 0, 1, 1)$$

وتعتبر عملية اختيار مجموعة السّمات من أهم العمليات في هذا النموذج، وذلك لأن فضاء الأشعة ذو بعد مساوٍ لعدد السّمات في هذه المجموعة وبالتالي زيادة عدد السّمات ستؤدي إلى زيادة في زمن التنفيذ. وعادة ما يتم اختيار مجموعة من الكلمات بحسب المجال المعمول عليه والتي تعبر أكثر ما يمكن عن هذا المجال.

ولما كان NMPlagiarism غير محدد المجال، كان لابد من حساب هذه المجموعة من أجل كل مجموعة من الملفات المراد مقارنتها وذلك بافتراض أن هذه المجموعة تتحدث في نفس المجال، وهو ما توفره لنا أدوات البحث المستخدمة لتحديد الملفات المراد استخدامها في المقارنة.

يقوم NMPlagiarism بإيجاد التمثيل الشعاعي لكل فقرة في المحتوى النصي لكل ملف من الملفات المراد مقارنتها، ومن ثم يتم احتساب التشابه بين كل مقطع من الملف الأول مع كل مقطع من الملف الثاني وذلك بحساب Cosine الزاوية بين الشعاعين الممثلين لكل من المقطعين. ثم وبتعريف عتبة للتشابه يتم اعتبار كل مقطعين بدرجة تشابه أكبر من هذه العتبة متشابهين.

#### 4.4.2. خوارزميات المقارنة اعتماداً على المعنى (Semantic-Based Algorithms):

تتنتمي جميع خوارزميات المقارنة السابقة لعائلة الخوارزميات الإحصائية، وهي خوارزميات ذات أداء جيد في معظم الحالات التي يكون فيها الانتحال مقتصر على نسخ لجزء كما هو من المحتوى النصي للمصدر. ولكن ماذا لو قام المنتحل باستبدال بعض الكلمات في الجزء المنسوخ بمصادفات؟ تبدأ نتائج الخوارزميات الإحصائية بالتراجع كلما زادت عمليات الاستبدال بالمصادفات.

تقوم خوارزميات المقارنة اعتماداً على المعنى بدرجة التشابه في معنى النصوص المراد مقارنتها، وذلك باستخدام شبكة دلالية تمثل العلاقات بين معظم كلمات اللغة مثل الـ WordNet المذكورة سابقاً.

تعتبر نتائج هذه العائلة من خوارزميات المقارنة بين النصوص الأكثر دقة، إلا أن تعقيد هذه الخوارزميات كبير جداً وبالتالي تحتاج لزمن تنفيذ طويل لإنجاز عملية المقارنة، وهو ما يدفع بمطوري نظم معالجة اللغات الطبيعية لاستخدام الخوارزميات الإحصائية ما أمكن.

#### 4.4.2.1 التشابه الدلالي (Semantic Relatedness):

التشابه الدلالي هو درجة التشابه بين مفهومين في معجم ما. ولتحديد مكان التشابه – إن وجد – يعمل NMPiagiarism على تقسيم كل نص من النصوص المراد مقارنتها إلى جمل، وفق الطريقة المذكورة سابقاً. ومن ثم تحسب درجة تشابه كل جملة من الملف الأول مع كل جملة من الملف الثاني. ثم وبتعريف عتبة للتشابه يتم اعتبار كل جملتين بدرجة تشابه أكبر من هذه العتبة متشابهتين.

ولحساب درجة التشابه بين جملتين يتم احتساب درجة التشابه الدلالي لكل كلمة من الجملة الأولى مع كل كلمة من الجملة الثانية (وهو السبب الرئيسي في الزيادة الكبيرة في تعقيد هذه الخوارزميات)، ويتم بناء جدول بدرجات التشابه تلك. ومن ثم يُستخدم ذلك الجدول لحساب درجة التشابه بين الجملتين باستخدام افتراض (Heuristic) معين.

إلا أن العملية ليست بهذه السهولة! تتميز اللغات الطبيعية بشكل عام بغموضها ووجود معاني متعددة لكل كلمة فيها تقريباً. مثلاً كلمة Interest يمكن أن تحمل معنى فائدة من البنك Interest from a bank ويمكن أن تكون بمعنى اهتمام بشئ ما Interest in a subject. إذاً، في حال ورود كلمة Interest ضمن نص ما ستظهر حالة غموض ولذلك نحن بحاجة لطريقة لإزالة هذا الغموض. يوجد العديد من الطرق لإزالة الغموض منها: Michael Lesk Algorithm و The Adapted Michael Lesk Algorithm.

قبل البدء في شرح الخوارزميتين السابقتين سنقوم بوضع تعريف دقيق لمصطلح إزالة الغموض.

إزالة الغموض: هي عملية تحديد المعنى الأكثر ملائمة لكلمة تحمل أكثر من معنى ضمن جملة ما.

##### - The Lesk Algorithm:

تستخدم خوارزمية Lesk قاموس الكتروني (Machine readable dictionary)، لإزالة غموض كلمة متعددة المعاني (Polysemy). الفكرة بشكل عام كالتالي:

حساب العدد الأعظمي للكلمات المشتركة بين مسردين (Gloss)، حيث تزداد نسبة التشابه بازدياد عدد الكلمات المتشابهة.

##### ○ Algorithm:

- Retrieve from MRD (WordNet in our case) all sense definitions (Glosses) of the words to be disambiguated.
- Determine the definition overlap for all possible sense combinations.
- Choose senses that lead to highest overlap.

والمثال التقليدي لهذه الخوارزمية هو: Pine cone

##### ○ Pine:

- Kinds of evergreen tree with needle-shaped leaves.
- Waste away through sorrow or illness.

##### ○ Cone:

- Solid body which narrows to a point.
- Something of this shape whether solid or hollow.
- Fruit of certain evergreen trees.

$$\begin{aligned} \text{Pine\#1} \cap \text{Cone\#1} &= 0 \\ \text{Pine\#2} \cap \text{Cone\#1} &= 0 \\ \text{Pine\#1} \cap \text{Cone\#2} &= 1 \\ \text{Pine\#2} \cap \text{Cone\#2} &= 0 \\ \text{Pine\#1} \cap \text{Cone\#3} &= 2 \\ \text{Pine\#2} \cap \text{Cone\#3} &= 0 \end{aligned}$$

المشكلة في خوارزمية Lesk أنها لا تستفيد من المعاني التي قامت بإسنادها خلال عملية إزالة الغموض، حيث تقوم بإعادة العمليات من جديد بالنسبة لكل كلمة.

- The Adapted Micheal Lesk Algorithm:

تختلف خوارزمية Adapted Lesk عن الخوارزمية الأصلية في النقاط التالية:

- استخدام شبكة دلالية Semantic Net بدلاً من استخدام القاموس فقط. وبالتالي، أصبح بإمكاننا بالإضافة لإيجاد معنى الكلمة إيجاد معاني مرادفاتنا وكل الكلمات التي ترتبط بها بشكل عام.
- خوارزمية جديدة لحساب تقارب معاني المفردات (Measuring scoring overlap)، تعطي نتائج أدق من نتائج خوارزمية Lesk الأصلية.

مبدأ عمل الخوارزمية:

لإزالة غموض أي كلمة في جملة فيها N كلمة، نقوم بتطبيق الخطوات التالية على كل كلمة من كلمات الجملة، أي نقوم بتطبيقهم N مرة:

- 1- إذا كان N كبيراً جداً، نقوم باختيار K (k-nearest neighbor) كلمة حول الكلمة الهدف (الكلمة المراد إزالة غموضها)، وذلك لتسريع المعالجة. إذا كان  $k=4$ ، فسنددد كلمتين قبل الكلمة الهدف وكلمتين بعد الكلمة الهدف مع مراعاة الحالات الخاصة.
- 2- من أجل كل كلمة في السياق الذي تم اختياره في الخطوة 1، نقوم بإيجاد جميع المعاني الممكنة لها باعتبارها اسماً وفعلاً.
- 3- من أجل كل معنى (WordSense) من المعاني الناتجة عن الخطوة 2، نقوم بإيجاد التالي:
  - a. المسرد (Gloss) الخاص بالمعنى، والأمثلة المطروحة عن استخدامه في WordNet.
  - b. المسرد الخاص بمجموعة المترادفات (Synsets) المرتبطة به عن طريق علاقة التعميم (Hypernym). إذا وجد أكثر من hypernym لمعنى كلمة واحد تدمج جميع الـ Glosses لكل Hypernym في مسرد واحد.
  - c. المسرد الخاص بمجموعة المترادفات المرتبطة به عن طريق علاقة التخصيص Hyponym.
  - d. المسرد الخاص بمجموعة المترادفات المرتبطة به عن طريق علاقة Meronym.
  - e. المسرد الخاص بمجموعة المترادفات المرتبطة به عن طريق علاقة Troponym.
- 4- تشكيل جميع الأزواج الممكنة من المسارد الناتجة عن الخطوات السابقة، ثم حساب نسبة التشابه بين هذه الأزواج. ويكون الناتج النهائي:

$$\text{Overall score} = \sum \text{scores for each relation pair.}$$

في مثال Pine cone، يوجد لدينا 3 معاني لكلمة Pine و 6 معاني لكلمة Cone وبالتالي لدينا 18 زوج محتمل يجب اختيار واحد منهم فقط.

لحساب التشابه، أو بدقة أكثر التداخل Overlap، يتم استخدام تقنية تعمل على التمييز بين تداخل N كلمة منفصلة وتداخل N كلمة متتالية وبالتالي تعطي نتائج أدق. تعتمد هذه التقنية على Zipf's law الذي يفترض أن طول الكلمات يتناسب عكسياً مع استخدامها، الكلمات الأقصر هي الأكثر استخداماً والأطول هي الأقل استخداماً. يتم حساب التداخل بين سلسلتي محارف عن طريق إيجاد أطول سلسلة جزئية متتالية مشتركة بين السلسلتين مع عدد أعظمي من المحارف المتتالية The longest common sub-string with maximal consecutives. إذا كان لدينا تداخل بين سلسلتي محارف يتضمن N كلمة متتالية، فيكون وزن هذا التداخل  $N^2$ .

مثال: تداخل "ABC" له الوزن  $3^2=9$ ، بينما تداخلين منفصلين (غير متتاليين) "AB" و "C" يكون لهم الوزن التالي:  $22 + 11=5$ .

5- بعد إيجاد أوزان جميع المعاني، نختار المعنى ذا الوزن الأكبر. فيكون الأكثر ملائمة للكلمة الهدف ضمن السياق المحدد. وبذلك، يعطينا الخرج، بالإضافة إلى المعنى الأكثر ملائمة، الجزء من الحديث المرتبط بهذه الكلمة.

بعد الانتهاء من الخطوات السابقة، نكون قد أوجدنا المعنى الأكثر ملائمة لكل كلمة في جمل الدخل. الآن لحساب التشابه الدلالي بين هذه الجمل سنعتمد على التشابه الدلالي لكلمات هذه الجمل، ولحساب التشابه الدلالي بين الكلمات نقوم باستخدام *The path length-based similarity measurement*.

يعرف نوعين من العلاقات في WordNet:

- **الأول:** علاقات بين مجموعات المترادفات.
- **الثاني:** علاقات بين معاني الكلمات.

تدعى العلاقة بين مجموعتي مترادفات علاقة دلالية **Semantic relation** أما العلاقة بين معاني الكلمات فاسمها علاقة معجمية **Lexical relation**. في العلاقات المعجمية، تكون العلاقات بين عناصر مجموعات مترادفات مختلفة، بينما العلاقات الدلالية تكون بين المجموعات بحد ذاتها. مثال:

- علاقات دلالية: **hypernym, hyponym, holonym**
- علاقات معجمية: **antonym, the derived-form relation**
- **dark#n#1 is the antonym of light#n#10**
- **light#n#10 is the antonym of dark#n#1**
- The synset to which **light#n#10 belongs** is {**light#n#10, lighting#n#1**}
- **But lighting#n#1 is not an antonym of dark#n#1**; therefore, the antonym relation needs to be a lexical relation, not a semantic relation.

التشابه الدلالي هو حالة خاصة من الارتباط الدلالي حيث يتم التعبير عنه فقط بعلاقة IS-A.

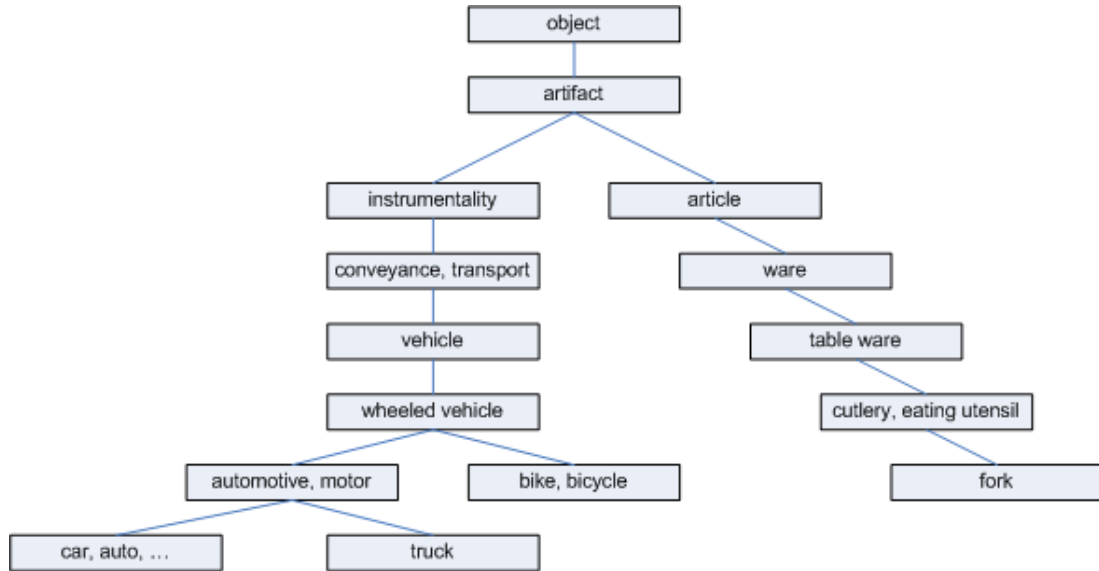
- **The path length-based similarity measurement:**

لحساب التشابه الدلالي بين مجموعتي مترادفات نقوم باستخدام العلاقات **hyponym/hypernym** (علاقات is-a) كما ذكرنا بالفقرة السابقة.

أسهل طريقة لحساب التشابه الدلالي بين مجموعتي مترادفات هو اعتبار تصنيف (taxonomy) الـ WordNet بيناناً غير موجهاً (undirected graph) ومن ثم حساب المسافة (المسار) بين المجموعتين. كلما كان المسار أقصر بين العقدتين كلما كانتا أكثر تشابهاً، حيث يقاس طول المسار بعدد العقد (Node) ويعتبر المسار بين عنصرين في مجموعة (synset) واحدة هو 1 (synonym relation).

يظهر الشكل التالي علاقات الـ Hyponym في WordNet





يمكن استخدام التصنيف في الشكل السابق لحساب التشابه بين الكلمات وذلك عن طريق حساب طول المسار بينهم مثلاً:

- طول المسار بين كلمة Car و Auto هو 1. (ضمن مجموعة واحدة)
- طول المسار بين كلمة Car و Truck هو 3. (عدد العقد أو المجموعات)
- طول المسار بين كلمة Car و Bike هو 4.
- طول المسار بين كلمة Car و Fork هو 12.

نطلق على أي عقدة (المجموعة) أب مشتركة بين مجموعتين Sub-sumer. ونطلق اسم Least common sub-sumer (LCS) على العقدة الأب المباشرة لعقدتين أخريين، بالعودة إلى المثال السابق:

LCS of {car, auto..} and {truck..} is {automotive, motor vehicle}

الآن بعد حساب طول المسار يجب حساب قيمة التشابه. تم طرح العديد من الطرق لحساب التشابه منها:

- $\text{Sim}(s, t) = 1/\text{distance}(s, t)$
- $\text{Sim}(s, t) = \text{SenseWeight}(s) * \text{SenseWeight}(t) / \text{PathLength}$ ; where:
  - s and t: denote the source and target words being compared.
  - SenseWeight: denotes a weight calculated according to the frequency of use of this sense and the total of frequency of use of all senses.
  - PathLength: denotes the length of the connection path from s to t.

الفصل الخامس:

# التصميم و التحقيق

5

## 5. التصميم و التحقيق

### 5.1 التصميم:

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

#### 5.1.1 الأجزاء الخارجية:

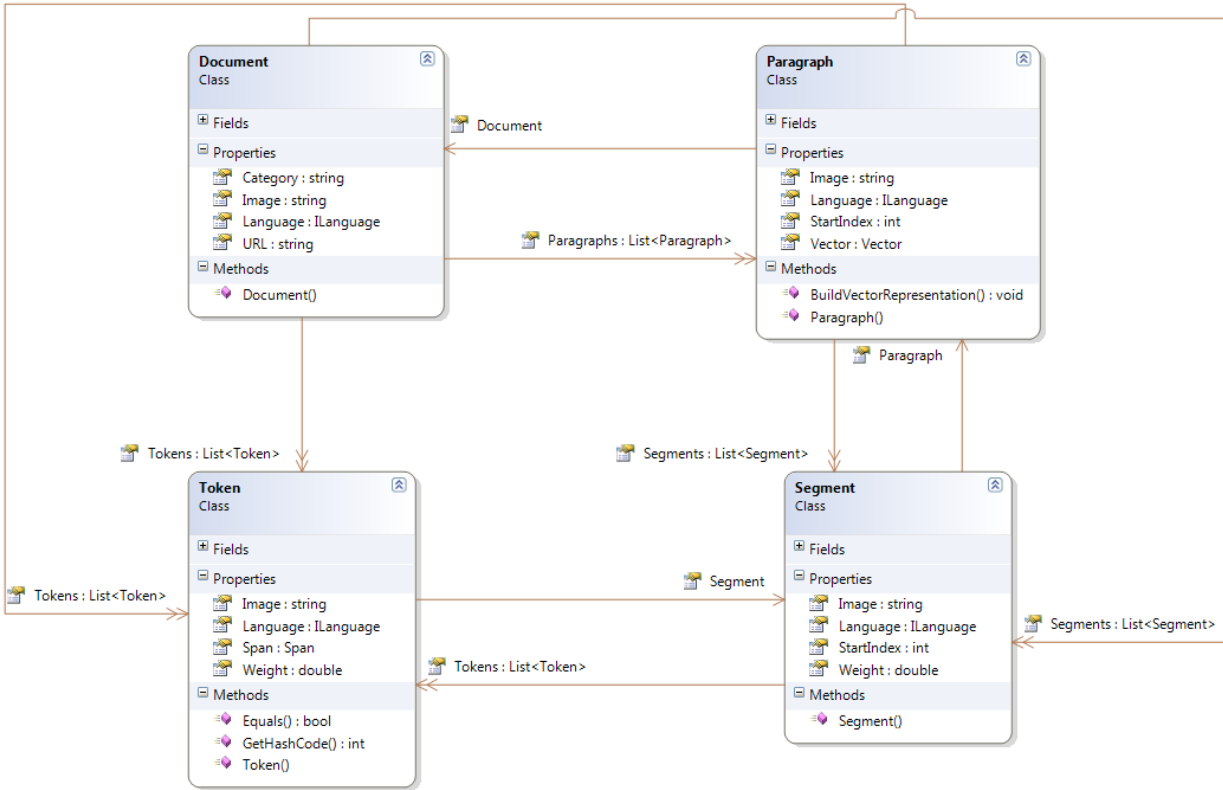
- OpenNLP: المحلل اللغوي.
- iTextSharp: مكتبة مجانية لقراءة الملفات من نمط PDF
- Majestic12: مكتبة مجانية و مفتوحة المصدر لقراءة صفحات HTML.

#### 5.1.2 الأجزاء الداخلية:

##### مخطط الصفوف (Class Diagram):

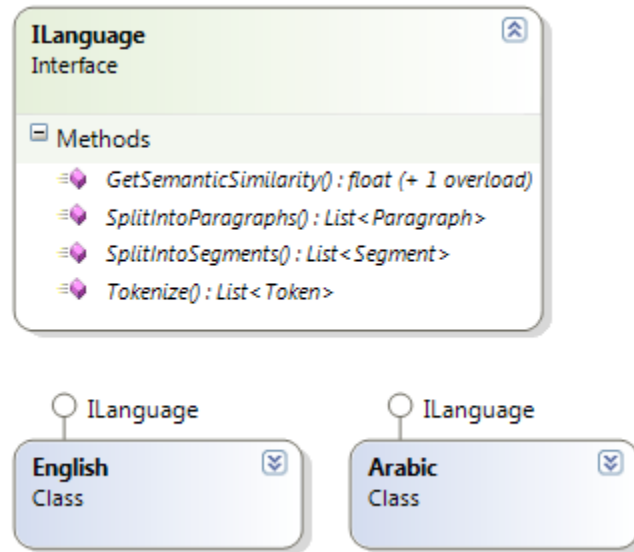
تم بناء NMPlagiarism كوحدات مستقلة فيما بينها، لكل منها مهمة معينة، و عُرِف لكل منها خرج محدد البنية ليكون دخلاً للوحدة التالية في مراحل عمل النظام. يتم التنسيق بين هذه الوحدات عن طريق وحدة رئيسية تمثل قلب النظام. إن هذا البناء يساعد في عملية توسيع أو تعديل بنية أي من وحدات NMPlagiarism دون التأثير في عمل باقي الوحدات طالما تمت المحافظة على بنية خرج الوحدة المعدلة. وعليه يمكن إضافة أية خوارزمية جديدة تحسن من نتائج NMPlagiarism بتحقيق هذه الخوارزمية بما يتناسب مع خرج الوحدة التي تنتمي إليها. وفيما يلي نبين مخطط الصفوف لكل وحدة من وحدات NMPlagiarism:

• الأغراض العامة (Utilities):



تدخل مكونات هذه الوحدة في عمل جميع وحدات النظام، ويتم فيها تمثيل النصوص الممثلة لدخل النظام وفق بنية محددة وواحدة أية كانت اللغة التي كتبت فيها هذه النصوص. حيث يمثل كل نص بغرض من نوع ملف نصي (*Document*) والذي يحوي بدوره قائمة من المقاطع النصية (*Paragraph*) وكل منها تحوي قائمة من الجمل (*Segment*) والتي بدورها تتألف من قائمة من الكلمات (*Token*).

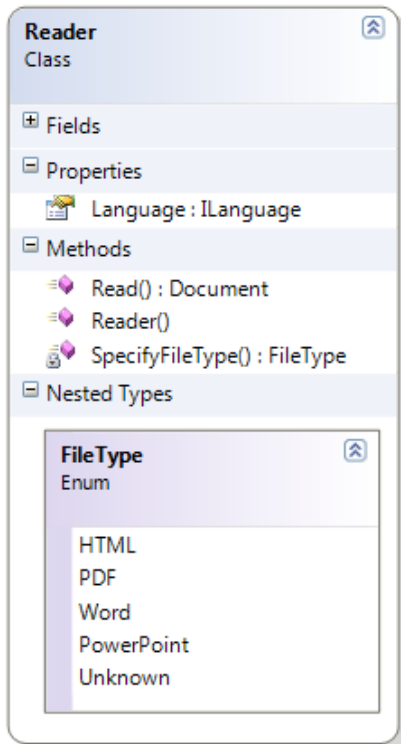
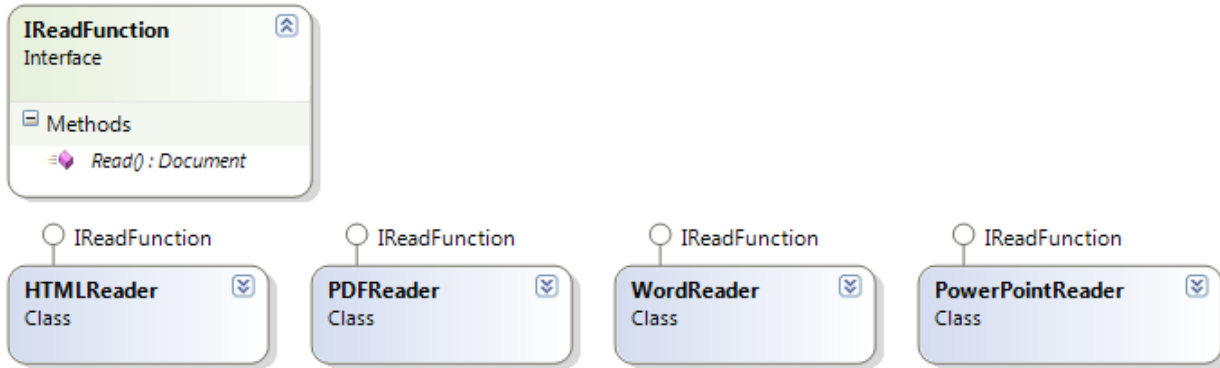
• اللغة (Language):



تؤثر هذه الوحدة في عمل جميع وحدات النظام لما لها من أهمية في معالجة النصوص الممثلة لدخل النظام، حيث تعمل هذه الوحدة على تقطيع النص إلى قائمة من المقاطع النصية وقائمة من الجمل وقائمة من الكلمات، ولكل من هذه القوائم استخدام في أداء مهمة ما. يتم هذا التقطيع لمرة واحدة من أجل كل نص جديد أثناء عمل النظام. وتستخدم هذه الوحدة في عملها مكتبات خارجية ومفتوحة المصدر لتقطيع النص إلى جمل وكلمات، في حين يتم التقطيع إلى مقاطع نصية باستخدام تجريبيات (Heuristic) خاصة بكل لغة. كما تستخدم شبكات دلالية خاصة بكل لغة من أجل حساب التشابه الدلالي بين كلمتين أو جملتين وذلك في حال تم استخدام أي من التوابع المعتمدة على معنى والمحققة في NMPlagiarism.

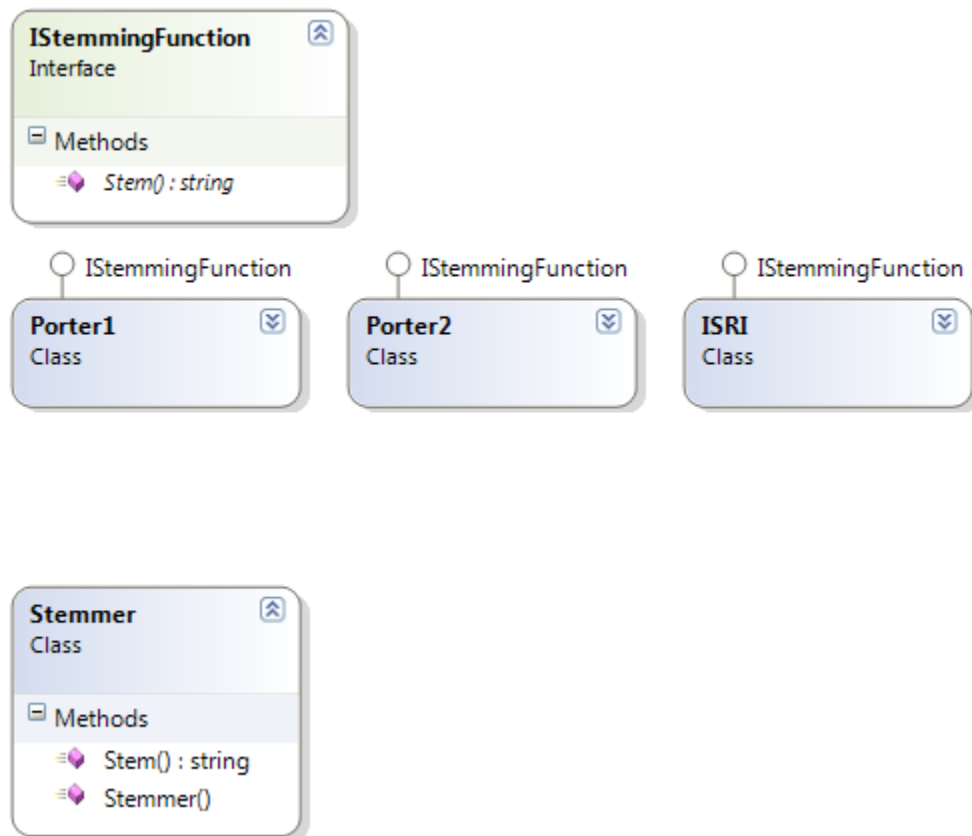
وكما توضح البنية المستخدمة يمكن إضافة أي لغة جديدة بتحقيق التوابع الأربعة المذكورة، وهذا مما يميز NMPlagiarism عن غيره من أنظمة كشف الانتحال حيث يمكن اعتباره مستقلاً عن اللغة (Language Independent). ولا يجب بالضرورة تحقيق التابع الأول (GetSemanticSimilarity()) في حال عدم توفر الشبكة الدلالية المناسبة للغة المراد إضافتها، ولكن وفي حال عدم تحقيق هذا التابع يجب ألا يتم استخدام أي من التوابع المعتمدة على المعنى في NMPlagiarism في حال كان الملف الدخول مكتوباً في اللغة الجديدة.

- قارئ الملفات (File Reader):



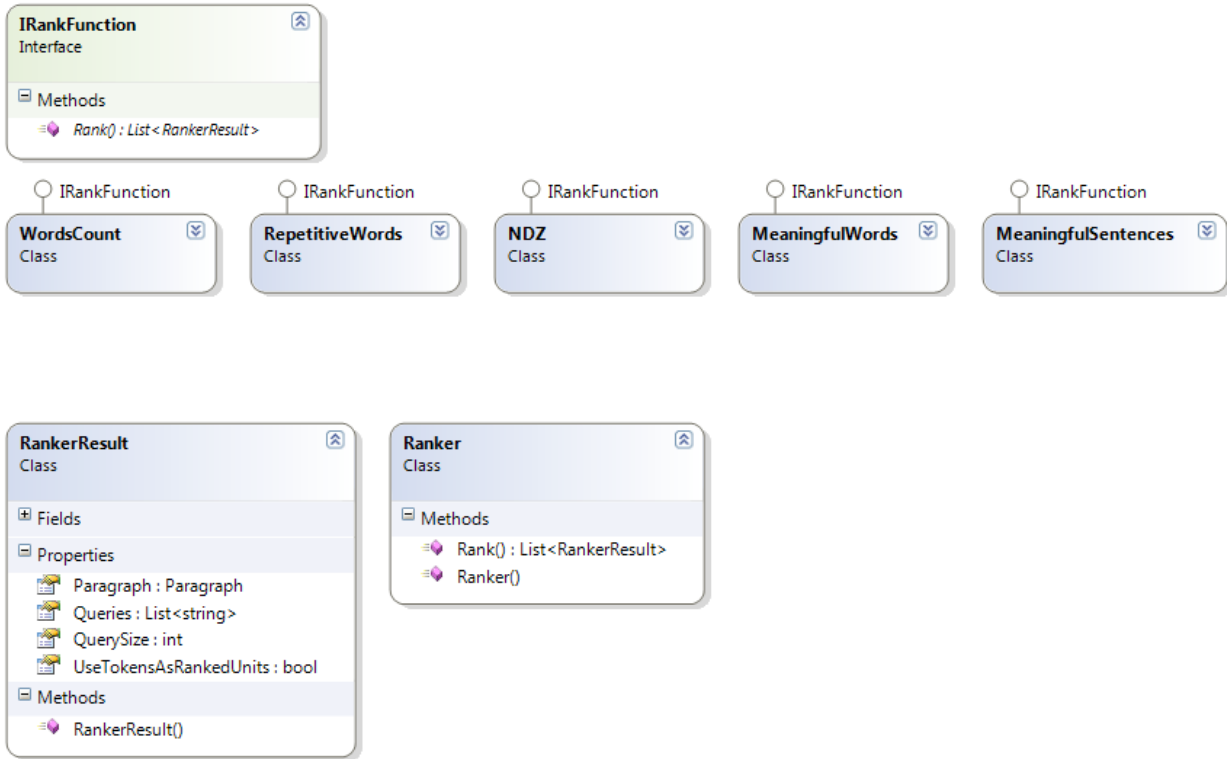
تعمل هذه الوحدة على قراءة الملفات النصية ذات الأنواع المدعومة في NMPlagiarism. ويتم في هذه الوحدة تحديد تلقائي لنوع الملف المراد قراءته واستخدام التابع المناسب لقراءة ذلك الملف حسب نوعه، وذلك باستخدام رابط الملف كدخل لهذه الوحدة سواء كان هذا الرابط لملف محلي أو ملف على الانترنت. يدعم NMPlagiarism أنواع الملفات النصية الشهيرة (HTML, PDF, Word, PowerPoint) والموضحة في الشكل.

- المجذع (Stemmer):



وتمثل الوحدة المسؤولة عن إيجاد جذع الكلمات. وقد تم تحقيق ثلاثة من أشهر توابع التجذيع في NMPPlagiarism، اثنان لتجذيع الكلمات في اللغة الإنكليزية (Porter1, Porter2)، وواحد لتجذع الكلمات في اللغة العربية (ISRI: The Arabic Stemmer of the Information Science Research Institute's Arabic Stemmer).

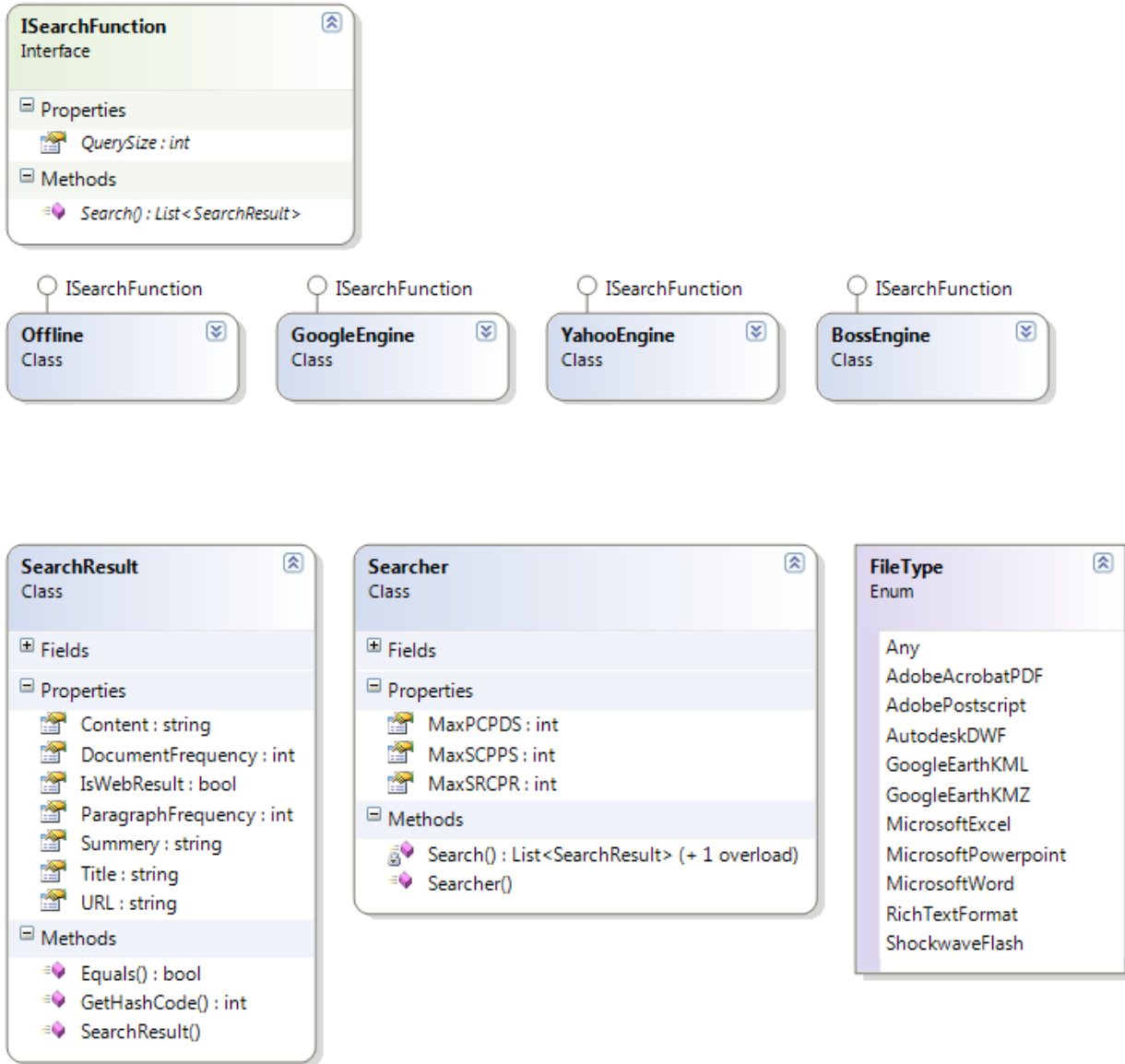
- مرتب الجمل (Ranker):



تعمل هذه الوحدة على تشكيل قائمة من الاستعلامات الممكنة لاستخدامها في عملية البحث مرتبة تنازلياً حسب أهمية هذه الاستعلامات. وذلك من أجل كل مقطع نصي (Paragraph) في الملف (Document) الممثل لدخل هذه الوحدة. ويتم تشكيل الاستعلامات باستخدام أحد توابع التوزيع المحققة في NMPlagiarism والموضحة في الشكل.



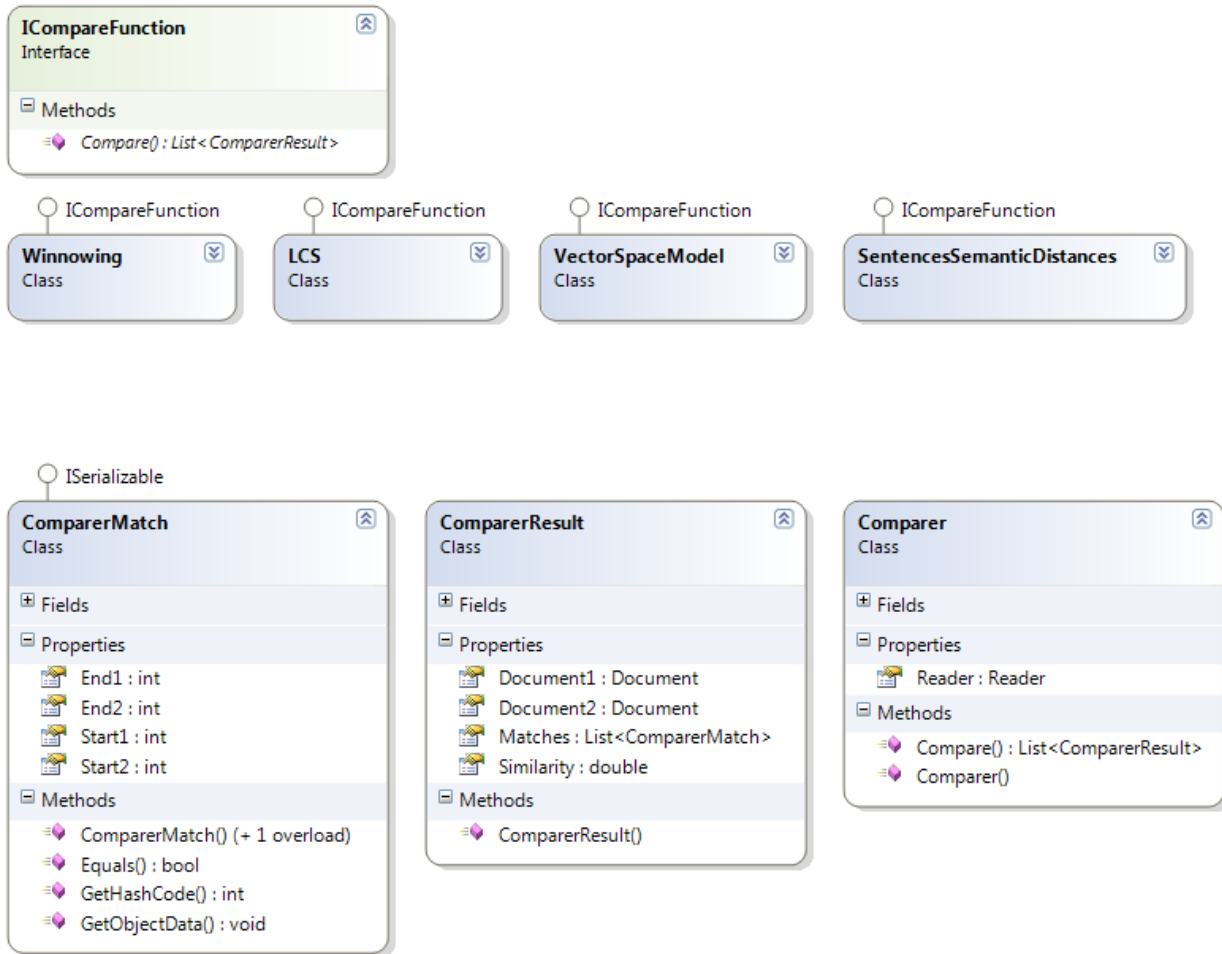
• الباحث (Searcher):



تستخدم هذه الوحدة خرج مرتب الجمل من الاستعلامات المرتبة حسب أهميتها في عملية البحث عن المصادر المتوقعة للملف الممثل لدخل النظام، وذلك باستخدام قاعدة معطيات محلية أو بالاستعانة بأحد محركات البحث الشهيرة ( Google, Yahoo, Boss ) كما هو موضح في مجموعة التوابع المحققة في NMPlagiarism والموضحة في الشكل.

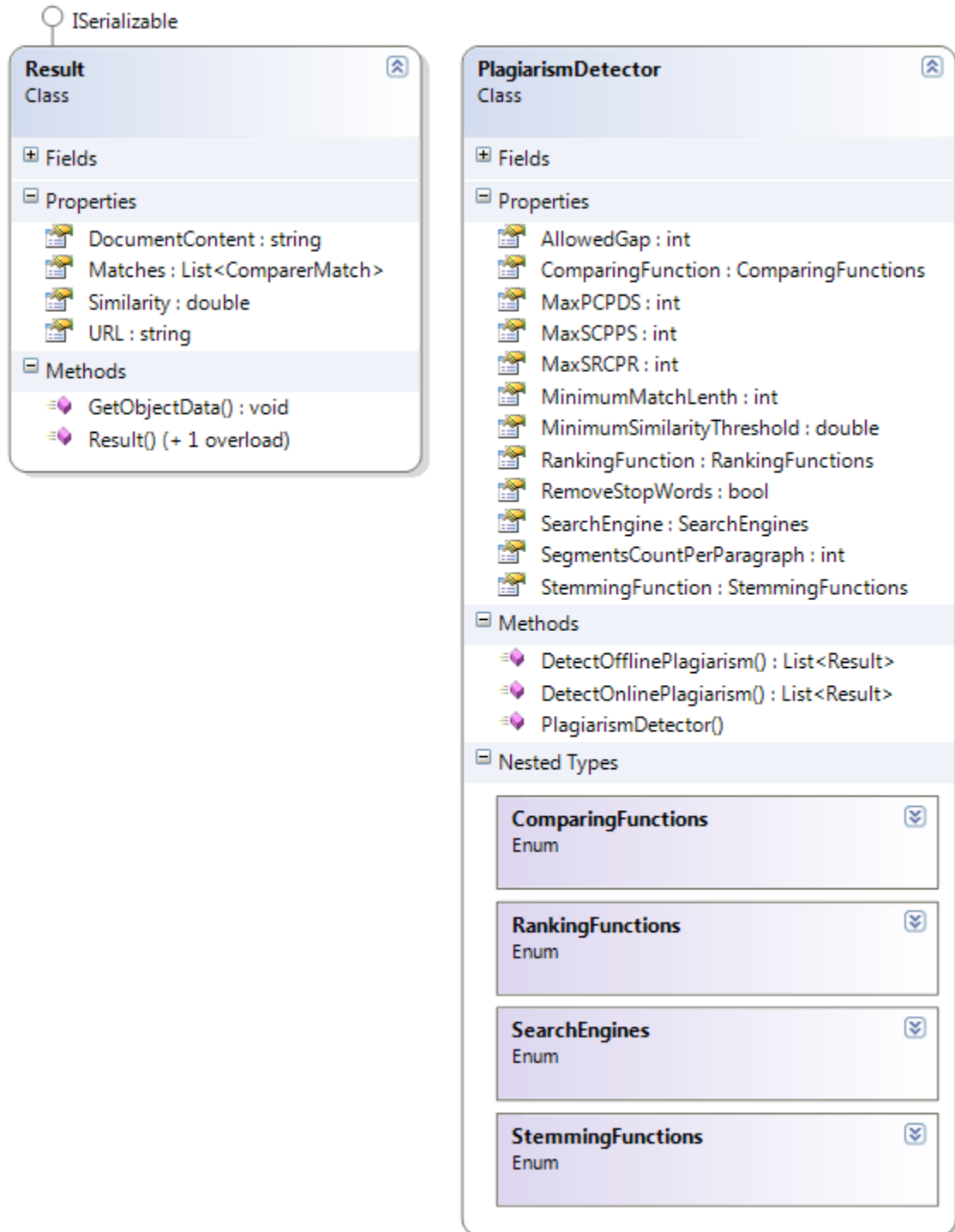
تعيد هذه الوحدة قائمة من نتائج البحث كل منها يحوي رابط لملف متوقع أن يكون مصدر للملف الدخلى ووزن يمثل أهمية هذه النتيجة على مستوى الملف (Document) بشكل عام ووزن لأهميتها على مستوى المقطع النصي (Paragraph) بشكل خاص، كما يمكن لنتيجة البحث أن تحوي المحتوى النصي للرابط في حال كان البحث قد تم ضمن قاعدة معطيات محلية.

• المقارن (Comparer):



مهمة هذه الوحدة والتي تعتبر من أهم وحدات النظام مقارنة المحتوى النصي للملف الدخل مع المحتويات النصية للملفات الموجودة ضمن نتائج البحث والممثلة لدخل هذه الوحدة، حيث وبلاستعانة بوحدة قراءة الملفات يتم قراءة المحتويات النصية لنتائج البحث غير المتضمنة لتلك المحتويات. ويمكن وكما هو موضح في الشكل استخدام واحد من عدة توابع مقارنة محققة في NMPPlagiarism، ثلاث منها إحصائية (Winnowing, LCS, VectorSpaceModel) وواحد بالاعتماد على المعنى (SentenceSemanticDistances).

- كاشف الانتحال (Plagiarism Detector):



تلخص هذه الوحدة مراحل عمل النظام وتشكل قلب هذا النظام حيث تعمل على التنسيق بين مختلف الوحدات بما يتناسب مع المهمة المطلوب تحقيقها، وذلك بالإشراف على تسليم كل وحدة الدخل الخاص بها وتسليم الخرج منها وفق البنى الموضحة لخرج كل وحدة من وحدات NMPlagiarism.

## 5.2 التحقيق:

هذا القسم يقدم نظرة سريعة على نتائج تحقيق البرنامج وبعض الميزات.  
من الجدير بالذكر أن هناك قسم للإعدادات يحوي خيارات التصدير والتصنيف والمستخدم في البرنامج كما سوف يتم التوضيح لاحقاً.

### your document

you can write, paste or upload your document, notice that the text that are in the following text box is the one that will be processed so make sure that it's the one you want to submit and if not you can modify it as you want before going on to with the step 1.

This study was conducted in a rural Pre-K through 8th grade school in northeastern Connecticut. Total enrollment is approximately 540 students. Graduating students attend a regional high school located in a neighboring community. There are 40 full-time classroom teachers on staff and 17 members of the support staff. A recent grant award designed to integrate technology within the social studies curriculum made it an optimal location to begin exploring the Internet searching skills the teachers possessed and how they acquired these skills.

Basic Search Strategy: The Ten Steps

The following list provides a guideline for you to follow in formulating search requests, viewing search results, and modifying search results. These procedures can be followed for virtually any search request, from the simplest to the most complicated. For some search requests, you may not want or need to go through a formal search strategy. If you want to save time in the long run, however, it's a good idea to follow a strategy, especially when you're new to a particular search engine.

### please specify some features which suit your search

النمط

Category

Statistical

محد ك البحث

Search Engine

Google

عدد الجمل في كل  
فقرة

Paragraph Sentences Number

5

44

## 5.2.1 اختيار إعدادات البرنامج:

### النص المدخل:

حيث يستطلع المستخدم وضع نصه الخاص. وبإمكانه رفعه كملف (MSWord, PDF, and HTML) ونص الملف سوف يتم عرضه. وبإمكان المستخدم تعديل النص المدخل.

### محرك البحث:

بإمكان المستخدم هنا اختيار محرك البحث بعد قاعدة البيانات المحلية. عدة خيارات متاحة:

- محرك بحث غوغل
- محرك بحث ياهو
- دون محرك بحث

### النمط:

هنا سوف يتم تحديد كيف يوف تتم عملية التصنيف.

إذا اختار المستخدم تصنيف إحصائي (Statistical) فسوف يتم اعتبار تابع الترتيب المختار من الإعدادات والتي تحوي تابع التوزيع الطبيعي (NDZ) كخيار افتراضي.

وإذا اختار المستخدم نمطا معيناً (غير التصنيف الإحصائي) فسوف يتم اعتبار تابع المسافة الدلالية مع النمط المزود به.

### عدد الجمل في كل فقرة:

يتم تقسيم النص المدخل إلى فقرات افتراضية. كل منها يحوي عدداً محدداً من الجمل. حيث يجب على المستخدم اختيار عدد الجمل في كل فقرة افتراضية. تابع التصنيف سوف يختار إحدى هذه الجمل في البحث.

وهذا العدد يجب أن يكون بين الواحد وعدد الجمل الكلية في النص. النظام يقوم بالتأكد من صلاحية القيمة المدخلة فيما إذا كانت تقع في المجال السابق.

هذه القيمة تؤثر على سرعة الأداء حيث القيم الصغيرة تؤدي إلى عدد فقرات افتراضية أكبر وبالتالي جمل أكثر للبحث عنها وبالتالي معالجة أبطأ. بينما القيم الكبيرة تؤدي إلى عدد فقرات وجمل أقل وبالتالي سوف تكون أسرع من السابقة.

### زر التقديم

عندما يختار المستخدم تقديم النص. يتم تقسيمه إلى فقرات افتراضية تتم معالجتها من قبل المرتب والباحث والمقارن.

## 5.2.2 إظهار النتائج:

# no more plagiarism

### النص الأصلي

This study was conducted in a rural Pre-K through 8th grade school in northeastern Connecticut. Total enrollment is approximately 540 students. Graduating students attend a regional high school located in a neighboring community. There are 40 full-time classroom teachers on staff and 17 members of the support staff. A recent grant award designed to integrate technology within the social studies curriculum made it an optimal location to begin exploring the Internet searching skills the teachers possessed and how they acquired these skills. Basic Search Strategy: The Ten Steps The following list provides a guideline for you to follow in formulating search requests, viewing search results, and modifying search results. These procedures can be followed for virtually any search request, from the simplest to the most complicated. For some search requests, you may not want or need to go through a formal search strategy. If you want to save time in the long run, however, it's a good idea to follow a strategy, especially when you're new to a particular search engine.

### النتائج

Similarity: 50.97312  
<http://www.webology.ir/2005/v2n1/a9.html>  
[ViewResult](#)

Similarity: 45.78313  
<http://www.webliminal.com/search/10steps.htm>  
[ViewResult](#)

Similarity: 34.10565  
<http://webliminal.com/search/search-web05.html>  
[ViewResult](#)

Similarity: 45.5051  
<http://webliminal.com/essentials/fyis/Cool/index.html>  
[ViewResult](#)

### النتائج:

وهنا يتم إظهار النتائج حيث تحوي كل نتيجة رابط إلى المقال الأصلي وزر إظهار النتيجة (view-results) لإظهار الأجزاء المشتركة بين النص المدخل والمقالات التي تم جلبها.

التشابه): رقم يعبر عن النسبة المئوية للتشابه بين النص المدخل والمقالات المجلوبة.

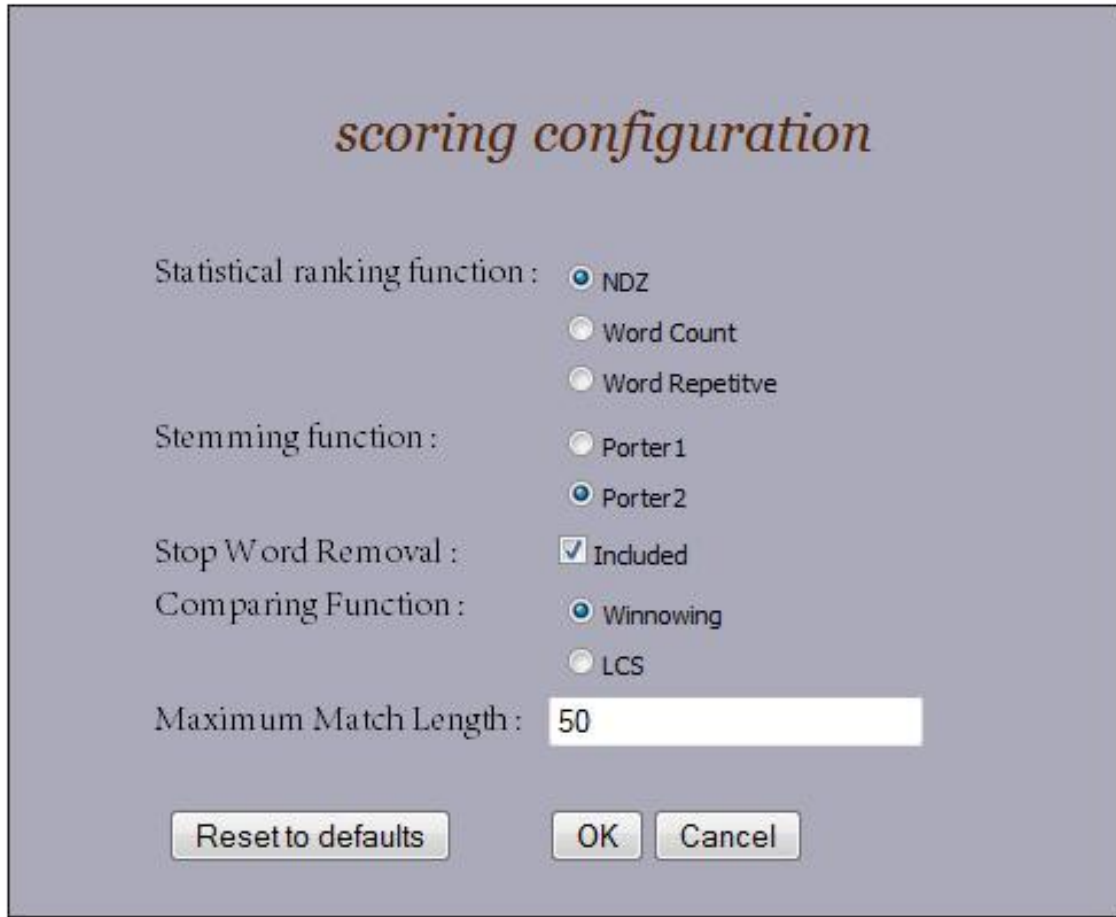
### النص الأصلي:

هنا يتم عرض وتلوين النص المدخل.

حيث الأجزاء ذات الخلفية الرمادية تعبر عن الجزء المنتحل بينما الأجزاء دون تلك الخلفية فهي من النص الأصلي.

### 5.2.3 الإعدادات:

تم وضع هذا القسم للمستخدمين ذوي خلفية جيدة حول كيفية عمل البرنامج.



The image shows a 'scoring configuration' dialog box with a light purple background. The title 'scoring configuration' is centered at the top in a brown, italicized font. Below the title, there are five configuration options, each with a label and a set of radio buttons or a checkbox. 1. 'Statistical ranking function :': Three radio buttons are shown: 'NDZ' (selected), 'Word Count', and 'Word Repetitive'. 2. 'Stemming function :': Two radio buttons are shown: 'Porter 1' and 'Porter 2' (selected). 3. 'Stop Word Removal :': A checkbox labeled 'Included' is checked. 4. 'Comparing Function :': Two radio buttons are shown: 'Winnowing' (selected) and 'LCS'. 5. 'Maximum Match Length :': A text input field containing the number '50'. At the bottom of the dialog, there are three buttons: 'Reset to defaults', 'OK', and 'Cancel'.

#### توابع الترتيب الإحصائية (Statistical Ranking Functions):

يمكن للمستخدم اختيار تابع التصنيف الذي يناسب نسه.

عدة خيارات متاحة أمام المستخدم:

- تابع عدد الكلمات (Word Count Function).
- تابع الكلمات المكررة (Repetitive word Function).
- تابع التوزع الطبيعي (NDZ Function).

الخيار الافتراضي هو تابع التوزع الطبيعي والذي تم اعتماده أدق التوابع بالتجريب.

#### توابع التصدير:

توابع التصدير المتاحة:

- Porter 1
- Porter 2

### حذف الكلمات الغير مفيدة:

عند تفعيل هذا الخيار سوف يتم تجاهل الكلمات الغير نفيدة خلال مرحلة البحث.

### توابع المقارنة:

التابع المختار سوف يتم استخدامه لعملية المقارنة بهدف الوصول للأجزاء المتشابهة بين النص المدخل والنص خرج مرحلة البحث. يمكن للمستخدم اختيار إحدى التوابع التالية:

- Winnowing

- LCS

### الطول الأصغري للتشابه:

هذا الرقم يحدد أقل عدد من المحارف المتسلسلة المشتركة ليتم اعتبار أن هناك تشابه بين النصين. قيم صغيرة سوف تؤدي إلى نتائج سيئة وغير متوقعة بينما قيم كبيرة لا تعطي أي تشابه.

## 5.3 التقنيات المستخدمة:

- Programming language: C#
- Microsoft .Net Framework 3.5
- LINQ to SQL
- ASP.NET
- Microsoft SQL Server 2008
- OpenNLP
- Word Net



## 5.4 مثال

يأخذ البرنامج كدخل نص أو رابط والذي تتم معالجته بواسطة القارئ حيث يكون الخرج على شكل نص. لنبدأ بالنص التالي والذي تم جمعه من ثلاث مواقع مختلفة ولنضعه كدخل للبرنامج.

### 5.4.1 الدخل:

*"Global warming is the increase in the average temperature of Earth's near-surface air and oceans since the mid-20th century and its projected continuation. According to the 2007 Fourth Assessment Report by the Intergovernmental Panel on Climate Change (IPCC), global surface temperature increased  $0.74 \pm 0.18$  C ( $1.33 \pm 0.32$  F) during the 20th century. Most of the observed temperature increase since the middle of the 20th century has been caused by increasing concentrations of greenhouse gases, which result from human activity such as the burning of fossil fuel and deforestation. Global dimming, a result of increasing concentrations of atmospheric aerosols that block sunlight from reaching the surface, has partially countered the effects of warming induced by greenhouse gases. "[from: [http://en.wikipedia.org/wiki/Global\\_warming](http://en.wikipedia.org/wiki/Global_warming)]*

*"Global warming is when the earth heats up (the temperature rises). It happens when greenhouse gases (carbon dioxide, water vapor, nitrous oxide, and methane) trap heat and light from the sun in the earth's atmosphere, which increases the temperature. This hurts many people, animals, and plants. Many cannot take the change, so they die. "[from: [http://library.thinkquest.org/CR0215471/global\\_warming.htm](http://library.thinkquest.org/CR0215471/global_warming.htm)]*

*"Global warming has become perhaps the most complicated issue facing world leaders. On the one hand, warnings from the scientific community are becoming louder, as an increasing body of science points to rising dangers from the ongoing buildup of human-related greenhouse gases — produced mainly by the burning of fossil fuels and forests. On the other, the technological, economic and political issues that have to be resolved before a concerted worldwide effort to reduce emissions can begin have gotten no simpler, particularly in the face of a global economic slowdown. "[from: <http://topics.nytimes.com/top/news/science/topics/globalwarming/index.html>]*

أولا تتم تجزئة النص إلى جمل باستخدام المكتبة المفتوحة المصدر (OpenNLP).

## 5.4.2 تجزئة النص إلى جمل:

Global warming is the increase in the average temperature of Earth's near-surface air and oceans since the mid-20th century and its projected continuation. According to the 2007 Fourth Assessment Report by the Intergovernmental Panel on Climate Change (IPCC), global surface temperature increased  $0.74 \pm 0.18$  C ( $1.33 \pm 0.32$  F) during the 20th century.

Most of the observed temperature increase since the middle of the 20th century has been caused by increasing concentrations of greenhouse gases, which result from human activity such as the burning of fossil fuel and deforestation.

Global dimming, a result of increasing concentrations of atmospheric aerosols that block sunlight from reaching the surface, has partially countered the effects of warming induced by greenhouse gases.

Global warming is when the earth heats up (the temperature rises). It happens when greenhouse gases (carbon dioxide, water vapor, nitrous oxide, and methane) trap heat and light from the sun in the earth's atmosphere, which increases the temperature.

This hurts many people, animals, and plants.

Many cannot take the change, so they die.

Global warming has become perhaps the most complicated issue facing world leaders.

On the one hand, warnings from the scientific community are becoming louder, as an increasing body of science points to rising dangers from the ongoing buildup of human-related greenhouse gases — produced mainly by the burning of fossil fuels and forests. On the other, the technological, economic and political issues that have to be resolved before a concerted worldwide effort to reduce emissions can begin have gotten no simpler, particularly in the face of a global economic slowdown.

ثم كل مجموعة من الجمل يتم تجميعها مرة أخرى لتكون فقرات افتراضية ليتم البحث من خلالها. وعدد الجمل في الفقرات الافتراضية يتم تحديده من المستخدم.

بفرض كل أربع جمل تكون فقرة.

### 5.4.3 تجميع الجمل ضمن فقرات:

Global warming is the increase in the average temperature of Earth 's near-surface air and oceans since the mid-20th century and its projected continuation .According to the 2007 Fourth Assessment Report by the Intergovernmental Panel on Climate Change ( IPCC ) , global surface temperature increased  $0.74 \pm 0.18$  C (  $1.33 \pm 0.32$  F ) during the 20th century .Most of the observed temperature increase since the middle of the 20th century has been caused by increasing concentrations of greenhouse gases , which result from human activity such as the burning of fossil fuel and deforestation .Global dim ming , a result of increasing concentrations of atmospheric aerosols that block sunlight from reaching the surface , has partially countered the effects of warming induced by greenhouse gases .

Global warming is when the earth heats up ( the temperature rises ) .It happens when greenhouse gases ( carbon dioxide , water vapor , nitrous oxide , and methane ) trap heat and light from the sun in the earth's atmosphere , which increases the temperature .This hurts many people , animals , and plants .Many cannot take the change , so they die .

Global warming has become perhaps the most complicated issue facing world leaders .On the one hand , warnings from the scientific community are becoming louder , as an increasing body of science points to rising dangers from the ongoing buildup of human-related greenhouse gases — produced mainly by the burning of fossil fuels and forests .On the other , the technological , economic and political issues that have to be resolved before a concerted worldwide effort to reduce emissions can begin have gotten no simpler , particularly in the face of a global economic slowdown .

ثم يتم تصنيف وترتيب جمل كل فقرة افتراضية اعتمادا على أهميتهم ضمن فقراتهم وذلك باستخدام المرتب.

باستخدام تابع الاصطفاء الطبيعي NDZ التابع للمرتب يتم تصنيف وترتيب جمل الفقرة السابقة كما يلي.

#### 5.4.4 تصنيف وترتيب الجمل ضمن كل فقرة افتراضية:

Global warming is the increase in the average temperature of Earth 's near-surface air and oceans since the mid-20th century and its projected continuation .  
Most of the observed temperature increase since the middle of the 20th century has been caused by increasing concentrations of greenhouse gases , which result from human activity such as the burning of fossil fuel and deforestation .  
Global dim ming , a result of increasing concentrations of atmospheric aerosols that block sunlight from reaching the surface , has partially countered the effects of warming induced by greenhouse gases .  
According to the 2007 Fourth Assessment Report by the Intergovernmental Panel on Climate Change ( IPCC ) , global surface temperature increased  $0.74 \pm 0.18$  C (  $1.33 \pm 0.32$  F ) during the 20th century .

Global warming is when the earth heats up ( the temperature rises ) .  
It happens when greenhouse gases ( carbon dioxide , water vapor , nitrous oxide , and methane ) trap heat and light from the sun in the earth's atmosphere , which increases the temperature .  
This hurts many people , animals , and plants .  
Many cannot take the change , so they die .

On the one hand , warnings from the scientific community are becoming louder , as an increasing body of science points to rising dangers from the ongoing buildup of human-related greenhouse gases — produced mainly by the burning of fossil fuels and forests .  
On the other , the technological , economic and political issues that have to be resolved before a concerted worldwide effort to reduce emissions can begin have gotten no simpler , particularly in the face of a global economic slowdown .  
Global warming has become perhaps the most complicated issue facing world leaders .

أهم جملة ضمن كل فقرة افتراضية يتم استخدامها للبحث نيابة عن فقرتها باستخدام قاعدة بيانات محلية أو باستخدام الانترنت.

باستخدام محرك البحث غوغل وباختيار خمس نتائج بحث تكون نتائج البرنامج كالتالي.

#### 5.4.5 البحث عن أهم جملة ضمن كل فقرة افتراضية:

[http://en.wikipedia.org/wiki/Global\\_warming](http://en.wikipedia.org/wiki/Global_warming)  
<http://www.recon2020-movie.com/global-warming.htm>  
<http://www.facebook.com/group.php?gid=20407481089>  
<http://www.facebook.com/group.php?gid=24505593707>  
<http://answers.yahoo.com/question/index?qid=20100514194926AA8jM5p>

---

[http://library.thinkquest.org/CR0215471/global\\_warming.htm](http://library.thinkquest.org/CR0215471/global_warming.htm)  
[http://en.wikipedia.org/wiki/Global\\_warming](http://en.wikipedia.org/wiki/Global_warming)  
[http://www.ecn.ac.uk/Education/climate\\_change.htm](http://www.ecn.ac.uk/Education/climate_change.htm)  
<http://ngm.nationalgeographic.com/ngm/0409/feature1/>  
<http://www.npr.org/templates/story/story.php?storyId=88520025>

---

<http://topics.nytimes.com/top/news/science/topics/globalwarming/index.html>  
<http://www.greenearthbaking.com/forums/how-do-you-practice-good-business/114>  
<http://flagcounter.boardhost.com/viewtopic.php?id=3090>  
<http://www.facebook.com/pages/payal-Mittal-Bharti/123099201061507>  
<http://en.rian.ru/Environment/20100706/159713252.html>

بعد حذف التكرارات في نتائج بحث الفقرات الافتراضية، كل نتيجة تتم قراءتها من قبل قارئ الملفات. وأخيرا تتم مقارنة خرج قارئ الملفات مع النص الأصلي وتكون النتائج هي خرج البرنامج والتي يتم تمثيلها بيانيا.

الفصل السادس:

# الاختبارات والنتائج

6

## 6. الاختبارات والنتائج

في هذا القسم تتم مناقشة نتائج البرنامج. التجارب موضحة في الملحق ب.

### توابع الترتيب:

تم تجريب توابع الترتيب على مجموعة من المفات. وأظهرت النتائج مايلي:

- عندما تكون الجملة المقترحة للبحث طويلة جميع التوابع تعطي نفس الفعالية.
- لأجل جمل قصيرة يعطي تابع التوزيع الطبيعي (NDZ function) نتائج أفضل.
- وقت التنفيذ: جميع توابع الترتيب تعكس نفس القيمة لأنها جميعها تعتمد على تكرارات الكلمات.
- لأجل بعض الفقرات الافتراضية جميع التوابع تفشل في استخراج الجملة المناسبة المعبرة عن الفقرة وهذا يؤدي إلى عدم إظهار النص المنتحل.

### محركات البحث:

- محرك البحث غوغل يستطيع جلب وثائق أحدث مقارنة بمحرك البحث ياهو.
- تختلف نتائج محركات البحث لأجل نفس التساؤل بين محرك البحث الأساسي وبين استخدام واجهته التطبيقية (API).
- بشكل عام فإن محرك البحث غوغل يعطي نتائج أفضل من محرك البحث ياهو.

### خوارزميات المقارنة:

LCS	Winnowing	
1050	1050	طول النص الأصلي
834	834	طول النص المنتحل
138	72	زمن المقارنة
1.75%	0.7%	نسبة النتيجة السالبة الخاطئة <sup>1</sup>
6%	6%	نسبة النتيجة الموجبة الخاطئة <sup>2</sup>

- لأجل فقرات افتراضية قصيرة لافرق يحدث.
- فقرت افتراضية كبيرة تجعل خوارزمية LCS تقطع بعض المحارف وبعض الأحيان بضع كلمات حيث خوارزمية Winnowing تقلل من قطع الكلمات والمحارف.
- معايير الأداء:
  - o LCS: يمكن أن تعتمد على الكلمات أو المحارف وهذه الأخيرة تعطي خيارات أخرى لتحديد التشابه.
  - o Winnowing: فقط اعتمادا على المحارف وأسرع وموثوقية أعلى.

<sup>1</sup> السالبة الخاطئة: النص غير منتحل (سالبة) ولكن النتيجة تظهر أنه منتحل (نتيجة خاطئة).

<sup>2</sup> الموجبة الخاطئة: النص منتحل (موجب) ولكن النتيجة تظهر أنه غير منتحل (نتيجة خاطئة).

### فحص النظام عامة:

- عدد الجمل في الفقرات الافتراضية حساس جدا حيث القيم الصغيرة تعطي نتائج إيجابية خاطئة كثيرة.
- هذا الرقم سوف يؤثر أيضا على زمن التنفيذ حيث يزداد زمن المعالجة لأجل القيم الصغيرة.
- عندما يتم اختيار عدد الجمل في الفقرات الافتراضية بشكل مناسب بعض المحارف لا تظهر في الخرج وهذه الأخطاء تحدث خلال عملية المقارنة.

### محرك البحث غوغل:

لا يمكن استخدام واجهة بحث غوغل التطبيقية لأجل البحث الاتوماتيكي.

المشكلة الأساسية عند استخدام IP ثابت (البرنامج عند المضيف) حيث تمنع غوغل خدماتها وبالتالي يجب أن يكون IP ديناميكي.

### محرك البحث BOSS:

سواء ثابت أو ديناميكي (IP). يمكنها التعامل مع أي Build your Own Search Service متوفرة من ياهو وهي اختصار لـ ولكن نتائج البحث تختلف عن غوغل.



# الفصل السابع:

## الخلاصة والآفاق المستقبلية

7

## 7. الخلاصة و الآفاق المستقبلية:

### 7.1 المشاكل

#### 7.1.1 مشاكل Google API:

نظراً لأن شركة Google لا تسمح باستخدام محرك بحثها من أجل البحث التلقائي (Automated search) فتتم 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).

#### 7.1.2 مشاكل 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 لـ Yahoo. وهذا موضح في الاختبارات.

### 7.2 الخلاصة:

ناقشنا في هذا التقرير نظاماً لكشف الانتحال في نصوص اللغات الطبيعية. تميز النظام عن الأنظمة الأخرى بخاصية الاختيار التلقائي للجمال التي يجب البحث عنها على الانترنت حيث اعتمدنا على مبدأ توزيع 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. أما 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 فحص 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.

### 7.3 الآفاق المستقبلية:

ملحق آ:

المراجع

آ

- [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-GuangTeng. (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] ZdeněkČeška 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](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 ChudaaandPavolNavrata. (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\\_\(computing\)](http://en.wikipedia.org/wiki/Fingerprint_(computing))
- [13] Saul Schleimer, Daniel Shawcross Wilkerson, Alexander Aiken. "Winnowing: Local algorithms for document fingerprinting". In *Proceedings of SIGMOD Conference'2003*.pp.76~85.
- [14] *Longest common subsequence problem*. Retrieved  
fromWikipedia.org:[http://en.wikipedia.org/wiki/Longest\\_common\\_subsequence\\_problem#cite\\_note-0](http://en.wikipedia.org/wiki/Longest_common_subsequence_problem#cite_note-0)
- [15] *Longest Common Subsequence*. Retrieved fromAlgorithmist.com  
[http://www.algorithmist.com/index.php/Longest\\_Common\\_Subsequence](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):  
<http://opennlp.sourceforge.net/>
- [18] *OpenNLP Projects*. Retrieved from [opennlp.sourceforge.net](http://opennlp.sourceforge.net):  
<http://opennlp.sourceforge.net/projects.html>
- [19] MSDN Library, Full-Text Search Concepts.
- [20] *Stemming*. Retrieved from Wikipedia.org:  
<http://en.wikipedia.org/wiki/Stemming>
- [21] <http://tartarus.org/~martin/PorterStemmer/>
- [22] Maurer, H., F. Kappe, B. Zaka. Plagiarism – A Survey. *Journal of Universal Computer Sciences*, vol. 12, no. 8, pp. 1050 – 1084, 2006.
- [23] Eissen, S., and Stein, B. Intrinsic Plagiarism Detection. Springer-Verlag ECIR LNCS 3936, pp. 565–569, 2006.
- [24] Gruner, S., S. Naven. Tool support for plagiarism detection in text documents. *Proceedings of the 2005 ACM Symposium on Applied Computing*. pp. 776 – 781, 2005.

## ملحق ب : الاختبارات

### 6.1 اختبارات 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

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
لم تظهر	[ 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 .	<b>Word Count Function</b>
النتيجة الأولى	A game engine is a software system designed for the creation and development of video games.	<b>Repetitive Word Function</b>
النتيجة الأولى	A game engine is a software system designed for the creation and development of video games.	<b>NDZ Function</b>

#### الاختبار الثاني :

الملف: On Automatic Plagiarism Detection

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	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	<b>Word Count Function</b>
النتيجة الأولى	On Automatic Plagiarism Detection Based on n-Grams Comparison Abstract .	<b>Repetitive Word Function</b>
النتيجة الأولى	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 their potential source .	<b>NDZ Function</b>

### 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 .	<b>Word Count Function</b>
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 .	<b>Repetitive Word Function</b>
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 .	<b>NDZ Function</b>

### 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 .	<b>Word Count Function</b>
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	<b>Repetitive Word Function</b>
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	<b>NDZ Function</b>

#### الاختبار الخامس :

الملف : A Web-Enabled Plagiarism Detection Tool

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	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 .	<b>Word Count Function</b>
النتيجة الأولى	A Web-Enabled Plagiarism Detection Tool This material is presented to ensure timely dissemination of scholarly and technical work .	<b>Repetitive Word Function</b>
النتيجة الأولى	A Web-Enabled Plagiarism Detection Tool This material is presented to ensure timely dissemination of scholarly and technical work .	<b>NDZ Function</b>

#### الاختبار السادس :

الملف : Extending Web Search for Online Plagiarism Detection

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	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 .	<b>Word Count Function</b>
النتيجة الأولى	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.	<b>Repetitive Word Function</b>
النتيجة الخامسة	To solve the data overloading problem, people commonly use web search engines to find what they need .	<b>NDZ Function</b>



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 .	<b>Word Count Function</b>
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 .	<b>Repetitive Word Function</b>
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 .	<b>NDZ Function</b>

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 .	<b>Word Count Function</b>
النتيجة الأولى	Plagiarism detection using software tools : a study in a Computer Science degree Keywords Plagiarism prevention and detection , e-learning , e-evaluation .	<b>Repetitive Word Function</b>
النتيجة الأولى	Plagiarism presents particular features in Computer Science and related degrees , such as ... sharing of knowledge has to be promoted among students.	<b>NDZ Function</b>

الملف : SNITCH A Software Tool for Detecting

ترتيب الملف الأصلي عند البحث في بواسطة Google	الجملة	التابع
النتيجة الأولى	In general , the algorithm uses the following steps : □ Open a document □ Analyze the document ... and other pertinent statistics .	<b>Word Count Function</b>
النتيجة الأولى	SNITCH : A Software Tool for Detecting Cut and Paste Plagiarism ABSTRACT Plagiarism of material from the Internet is a widespread and growing problem .	<b>Repetitive Word Function</b>
النتيجة الأولى	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 expository than in liberal arts courses .	<b>NDZ Function</b>

## 8.1 اختبارات واجهات محركات البحث:

تم تنفيذ هذا النوع من الاختبارات لكي نستطيع تحديد أي محرك بحث أفضل بالنسبة للنظام. يتم الاختبار بإرسال 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 .

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

#### الاختبار الخامس :

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

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

#### 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 .

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

#### 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 .

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

## 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 تبين 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](http://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](http://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.

## 8.2 اختبارات خوارزميات المقارنة:

استخدمنا في المقارنة المصطلحات التالية للتعبير عن سير البرنامج :

**الإيجابية الخاطئة** : توجد عندما يكون النص غير منتهل (سلبية) ويظهر البرنامج أن النص منتهل (نتيجة خاطئة) وعبرنا عن هذه الخاصة **باللون الأخضر**.

**السلبية الخاطئة** : توجد عندما يكون النص منتهل (إيجابية) ويظهر البرنامج أن النص غير منتهل (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 <i>seismograph</i> . 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 <i>seismograph</i> . 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:

نسبة السلبية الخاطئة	نسبة الإيجابية الخاطئة	زمن المقارنة (ميلي ثانية)	عدد حروف النص المقتبس	عدد حروف النص الأصلي	الخوارزمية
12 %	0.9 %	95	1156	1524	Winnowing
النص المقتبس			النص الأصلي		
<p>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. <b>The scale of the nucleation zone is uncertain, with some evidence.</b> Earthquake swarms are sequences of earthquakes striking in a specific area within a short period of time. <b>Once the rupture has initiated it begins to propagate along the fault surface.</b> Earthquakes often occur in volcanic regions and are caused there, both by tectonic faults and the movement of magma in volcanoes. <b>Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone.</b> Earthquake swarms can serve as markers for the location of the flowing magma throughout the volcanoes. <b>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.</b> These swarms can be recorded by seismometers and tiltmeters. <b>about a dozen earthquakes that struck the North Anatolian Fault in Turkey.</b></p>			<p>A tectonic earthquake begins by an initial rupture at a point on the fault surface, a process known as nucleation. <b>The scale of the nucleation zone is uncertain, with some evidence,</b> 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. <b>Once the rupture has initiated it begins to propagate along the fault surface.</b> <b>The mechanics of this process are poorly understood, partly because it is difficult to recreate the high sliding velocities in a laboratory.</b> <b>Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone.</b> 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. <b>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.</b> <b>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.</b></p>		

#### 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		
<p>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.</p>			<p>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.</p>		

#### النتائج:

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



### 8.3 اختبارات النظام كاملاً:

#### الاختبار الأول:

تابع انتقاء الجملة	محرك البحث	خوارزمية المقارنة	عدد حروف النص	الزمن (ثانية)	نسبة الإيجابية الخاطئة	نسبة السلبية الخاطئة	عدد الجمل في الفقرة
NDZ	Google	Winnowing	1394	119	0%	9%	14
النص							
<p>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.</p> <ol style="list-style-type: none"> <li>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.</li> <li>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.</li> <li>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.</li> <li>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.</li> <li>5. Blocked - A thread can enter in this state because of waiting the resources that are hold by another thread.</li> </ol>							
المواقع التي تم الاقتباس عنها :							
<a href="http://www.roseindia.net/java/thread/life-cycle-of-threads.shtml">http://www.roseindia.net/java/thread/life-cycle-of-threads.shtml</a>							

## 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							
<p>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</p> <p>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</p> <p>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</p>							
المواقع التي تم الاقتباس عنها :							
<a href="http://www.algorithmist.com/index.php/Dynamic_Programming">http://www.algorithmist.com/index.php/Dynamic_Programming</a> <a href="http://aymanh.com/syntax-highlighting-django-markdown-pygments">http://aymanh.com/syntax-highlighting-django-markdown-pygments</a>							

### الاختبار الثالث :

تابع انتقاء الجمل	محرك البحث	خوارزمية المقارنة	عدد حروف النص	الزمن (ثانية)	نسبة الإيجابية الصحيحة	نسبة السلبية الخاطئة	عدد الجمل في الفقرة
NDZ	Google	Winnowing	3101	529137.2649	0.11%	0%	4
النص							
<p>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.</p> <p>Maintenant, je crois que la profession qui je souhaite exercer est clair: Chercheur Chercheur, qu'est-ce que ça veut dire?</p> <p>1- Un chercheur (féminine chercheuse) une personne dont le métier consiste à faire de la recherche.</p> <p>2- Selon la définition de l'organisation de coopération et de développement économiques le chercheur est :</p> <p>« 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 »</p> <p>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?</p> <p>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....</p> <p>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.</p> <p>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.</p> <p>Les qualités nécessaires: Le chercheur doit être patient, compétent, persévérant et peut-être curieux.</p> <p>Les avantages: Intéressant, pas traditionnel, chaque jour il y a quelque chose nouvelle.</p> <p>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!</p> <p>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.</p> <p>Once you are given a problem, it is usually a good idea to check if DP is applicable to it.</p> <p>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.</p>							
المواقع التي تم الاقتباس عنها :							
<a href="http://aymanh.com/syntax-highlighting-django-markdown-pygments">http://aymanh.com/syntax-highlighting-django-markdown-pygments</a> <a href="http://fr.wikipedia.org/wiki/Chercheur">http://fr.wikipedia.org/wiki/Chercheur</a> <a href="http://www.algorithmist.com/index.php/Dynamic_Programming">http://www.algorithmist.com/index.php/Dynamic_Programming</a>							

## ملحق ج - WordNet:

WordNet: قاعدة بيانات معجمية مجانية متاحة على شبكة الإنترنت. صُمم الـ *WordNet* لإنشاء علاقات بين أربعة أنواع من أقسام الكلام (POS) – اسم Noun، فعل Verb، صفة Adjective وظرف Adverb. أصغر وحدة في الـ *WordNet* هي **synset** والتي تمثل مجموعة مرادفات.

تحتوي الـ **Synset** على الكلمة Word، تفسيرها Explanation، ومرادفاتها Synonyms. ويطلق على معنى محدد لكلمة واحدة ضمن قسم واحد من أقسام الكلام Sense.

فيما يلي شرح لمعنى العلاقات بين الأسماء والأفعال في WordNet:

- **Hypernyms:** *Y* is a hypernym of *X* if every *X* is a (kind of) *Y* (*canine* is a hypernym of *dog*, because every dog is a member of the larger category of canines).
- **Hyponyms:** *Y* is a hyponym of *X* if every *Y* is a (kind of) *X* (*dog* is a hyponym of *canine*).
- **Holonym:** *Y* is a holonym of *X* if *X* is a part of *Y* (*building* is a holonym of *window*).
- **Meronym:** *Y* is a meronym of *X* if *Y* is a part of *X* (*window* is a meronym of *building*).
- **Hypernym:** the verb *Y* is a hypernym of the verb *X* if the activity *X* is a (kind of) *Y* (*to perceive* is an hypernym of *to listen*).
- **Troponym:** the verb *Y* is a troponym of the verb *X* if the activity *Y* is doing *X* in some manner (*to lisp* is a troponym of *to talk*).