Algorithms for massive datasets project: Finding Similar Items on Stacksample

Letizia Molinari ID 959194

Abstract

The aim of the following project is the detection of pairs of textually similar documents in the Stacksample dataset. The discovery is performed by adopting the Locality Sensitive Hashing method for Jaccard distance.

1 Introduction

An important set of data-mining problems is the examination of data for 'similar' items. In the context of finding similar documents in a large corpus, which could be the Web or a collection of news articles, the character-level similarity could be useful for the detection of duplicates or near duplicates. The project below analyzes the StackSample dataset and implements a detector for the output of pairs of documents inferred as similar.

2 Dataset Description

The dataset Stacksample has been downloaded from Kaggle, where it is released, and it contains the text of the ten percent of the questions and answers from the website StackOverflow programming QA. The dataset is organized in three tables:

- the Questions table, which contains title, body, creation date, closed date(if applicable), score and owner Id for all non-deleted Stack Overflow questions whose Id is a multiple of 10,
- the Answers table, which contains body, creation date, score, and owner ID for each of the answers to the questions. A link back to the Questions table is created by the ParentId column,

- the Tags table, which contains the tags on each of the questions.

Only the Questions table is considered in this project and , in particular, the detector must only consider the 'Body' column of this table in order to output the pairs inferred as similar.

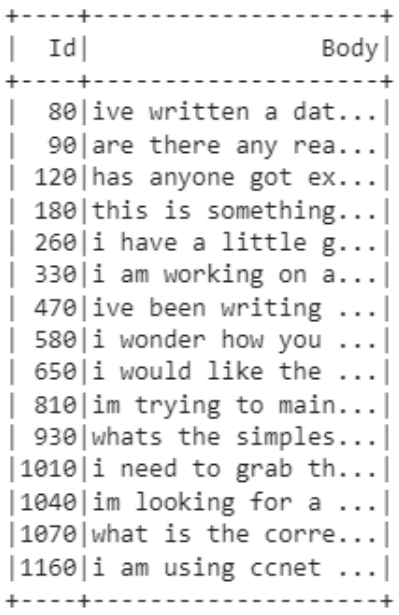
Id	OwnerUserId	CreationDate	ClosedDate	Score	Title	Body
80	26	2008-08-01T13:57:07Z	NA	26	SQLStatement.exec...	<p>I've written a...
90	58	2008-08-01T14:41:24Z	2012-12-26T03:45:49Z	144	Good branching an...	<p>Are there any ...
120	83	2008-08-01T15:50:08Z	NA	21	ASP.NET Site Maps	<p>Has anyone got...
180	2089740	2008-08-01T18:42:19Z	NA	53	Function for crea...	<p>This is someth...
260	91	2008-08-01T23:22:08Z	NA	49	Adding scripting ...	<p>I have a littl...
330	63	2008-08-02T02:51:36Z	NA	29	Should I use nest...	<p>I am working o...
470	71	2008-08-02T15:11:47Z	2016-03-26T05:23:29Z	13	Homegrown consump...	<p>I've been writ...
580	91	2008-08-02T23:30:59Z	NA	21	Deploying SQL Ser...	<p>I wonder how y...
650	143	2008-08-03T11:12:52Z	NA	79	Automatically upd...	<p>I would like t...
810	233	2008-08-03T20:35:01Z	NA	9	Visual Studio Set...	<p>I'm trying to ...
930	245	2008-08-04T00:47:25Z	NA	28	How do I connect ...	<p>What's the sim...
1010	67	2008-08-04T03:59:42Z	NA	14	How to get the va...	<p>I need to grab...
1040	254	2008-08-04T05:45:22Z	NA	42	How do I delete a...	<p>I'm looking fo...
1070	236	2008-08-04T07:34:44Z	NA	17	Process size on UNIX	<p>What is the co...
1160	120	2008-08-04T11:37:24Z	NA	36	Use SVN Revision ...	<p>I am using CCN...

Figure one: display of the Questions table

The table displayed in figure one reports the first fifteen rows in the questions table.

3 Dataset Preprocessing

In order to perform the analysis, first of all, only columns 'Id' and 'Body' of the Questions table have been selected, discarding all the other attributes in the table. Then, a function called cleaner() has been defined in order to remove from the text in the Body attribute all unwanted characteristics, such as quotes, urls, figures and apostrophes.



Id	Body
80	ive written a dat...
90	are there any rea...
120	has anyone got ex...
180	this is something...
260	i have a little g...
330	i am working on a...
470	ive been writing ...
580	i wonder how you ...
650	i would like the ...
810	im trying to main...
930	whats the simples...
1010	i need to grab th...
1040	im looking for a ...
1070	what is the corre...
1160	i am using ccnet ...

Figure two: cleaned text

The table in figure two reports the cleaned text. Then, the function RegexTokenizer in pyspark.ml.feature has been applied on the cleaned text in order to split it into individual units (the tokens). The tokens obtained in the previous step have been used as an input of the application of the StopWordsRemover function contained in pyspark.ml.feature. By applying this function, stop words have been removed, and it is possible to focus the analysis on the most relevant words in the text.

Id	Body	tokens	removed_stopwrd
80	ive written a dat...	[ive, written, a,...	[ive, written, da...
90	are there any rea...	[are, there, any,...	[really, good, tu...
120	has anyone got ex...	[has, anyone, got...	[anyone, got, exp...
180	this is something...	[this, is, someth...	[something, ive, ...
260	i have a little g...	[i, have, a, litt...	[little, game, wr...
330	i am working on a...	[i, am, working, ...	[working, collect...
470	ive been writing ...	[ive, been, writi...	[ive, writing, we...
580	i wonder how you ...	[i, wonder, how, ...	[wonder, guys, ma...
650	i would like the ...	[i, would, like, ...	[like, version, p...
810	im trying to main...	[im, trying, to, ...	[im, trying, main...
930	whats the simples...	[whats, the, simp...	[whats, simplest,...
1010	i need to grab th...	[i, need, to, gra...	[need, grab, base...
1040	im looking for a ...	[im, looking, for...	[im, looking, way...
1070	what is the corre...	[what, is, the, c...	[correct, way, ge...
1160	i am using ccnet ...	[i, am, using, cc...	[using, ccnet, sa...

Figure three: preprocessed dataset

The table in figure three reports the column 'tokens', which contains the tokens obtained from the original text, and the column removed-stopwrd, which contains the tokens in which stopwords have been removed.

4 Algorithms implementation

In the project, the following approaches have been applied:

- Local Sensitive Hash (LSH) for Minhash, which hashes items several times, in a way such that similar items are hashed to the same bucket with higher probability with respect to those which are dissimilar. Then any pair that hashed to the same bucket for any of the hashing is considered a candidate pair.

- Jaccard Distance, which is equal to 1 - Jaccard Similarity. The Jaccard similarity between two sets, called Y and Z, is the ratio between the size of the intersection between the two sets Y and Z and the size of the union of the two sets Y and Z. The Jaccard similarity is defined as follows:

$$Sim(Y, Z) = \frac{|Y \cap Z|}{|Y \cup Z|}$$

This means that the smaller the Jaccard distance is, the higher is the similarity between two documents.

First of all, the function HashingTF imported from pyspark.ml.feature has been implemented. This function maps a sequence of terms to their term frequencies using the hashing trick. It takes the text where stop words have been removed as input and it outputs a column of vectors called "TF". This column is used as input for the application of the function MinHashLSH, which outputs "hash", the hashes which will be used for the implementation of the Jaccard Distance.

Id	Body	tokens	removed_stopwrd	TF	hash
80	ive written a dat...	[ive, written, a,...	[ive, written, da...	(1024,[3,6,11,24,...	[[2078511.0]]
90	are there any rea...	[are, there, any,...	[really, good, tu...	(1024,[24,65,80,4...	[[5.527494E7]]
120	has anyone got ex...	[has, anyone, got...	[anyone, got, exp...	(1024,[19,39,88,9...	[[3.0469636E7]]
180	this is something...	[this, is, someth...	[something, ive, ...	(1024,[9,17,105,1...	[[2703620.0]]
260	i have a little g...	[i, have, a, litt...	[little, game, wr...	(1024,[24,43,53,5...	[[8.2415847E7]]
330	i am working on a...	[i, am, working, ...	[working, collect...	(1024,[79,108,117...	[[2703620.0]]
470	ive been writing ...	[ive, been, writi...	[ive, writing, we...	(1024,[29,58,70,1...	[[4.9813792E7]]
580	i wonder how you ...	[i, wonder, how, ...	[wonder, guys, ma...	(1024,[0,9,24,35,...	[[1.6586628E7]]
650	i would like the ...	[i, would, like, ...	[like, version, p...	(1024,[3,4,23,53,...	[[6.0110979E7]]
810	im trying to main...	[im, trying, to, ...	[im, trying, main...	(1024,[47,53,71,8...	[[2.2047776E7]]
930	whats the simples...	[whats, the, simp...	[whats, simplest,...	(1024,[58,153,168...	[[8.2415847E7]]
1010	i need to grab th...	[i, need, to, gra...	[need, grab, base...	(1024,[70,110,119...	[[5.8860761E7]]
1040	im looking for a ...	[im, looking, for...	[im, looking, way...	(1024,[31,81,88,1...	[[1.59167362E8]]
1070	what is the corre...	[what, is, the, c...	[correct, way, ge...	(1024,[271,353,37...	[[2.3404355E8]]
1160	i am using ccnet ...	[i, am, using, cc...	[using, ccnet, sa...	(1024,[3,17,65,71...	[[1.10806972E8]]

Figure four: application of Hashing TF and MinHashLSH functions

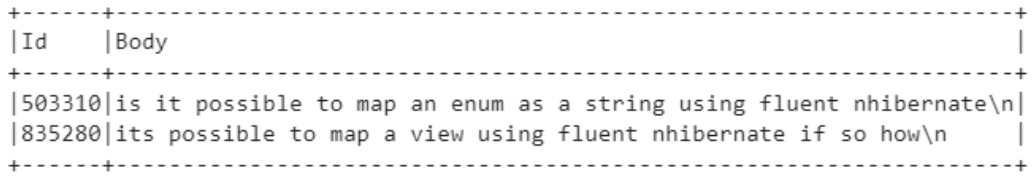
The table in figure four reports the two columns mentioned above, obtained after the application of the Spark functions. Then, all rows have been filtered in order to consider only rows which contain at least a word. In order to limit the computational time, a limit on the maximum number of rows has been set. In particular, it has been chosen to set this limit to 30000. The approximate similarity join has then been implemented, taking as input the hashed data. The dataset has been divided in two parts with respect to the ids, id-A and id-B, respectively and a threshold of 0.6 has been set. By doing so, it has been made possible to show items which have a similarity score below 0.6, in order to visualize only a suitable number of pairs.

5 Results

id_A	id_B	JaccardDistance
103560	1503630	0.5529411764705883
252660	920670	0.5
270440	272190	0.563636363636363
503310	835280	0.375
612820	634630	0.48
897770	905410	0.487179487179487
1041520	1042370	0.5
1071630	865480	0.5769230769230769
1082310	1276960	0.5642458100558659
1125640	198460	0.5714285714285714
1406050	936820	0.5906735751295337

Figure five: results

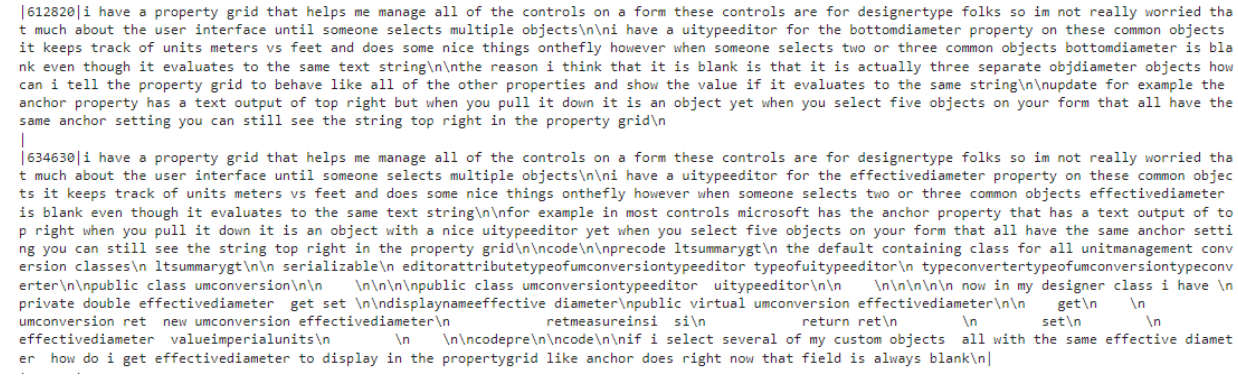
The table in figure five reports the couples which have a Jaccard Distance below 0.6. It is possible to observe that pairs which have the score closer to zero are the most similar ones. By looking at the table, it is possible to observe that questions having ids 503310 and 835280 are the most similar ones: they have a Jaccard Distance of 0.375.



Id	Body
503310	is it possible to map an enum as a string using fluent nhibernate\n
835280	its possible to map a view using fluent nhibernate if so how\n

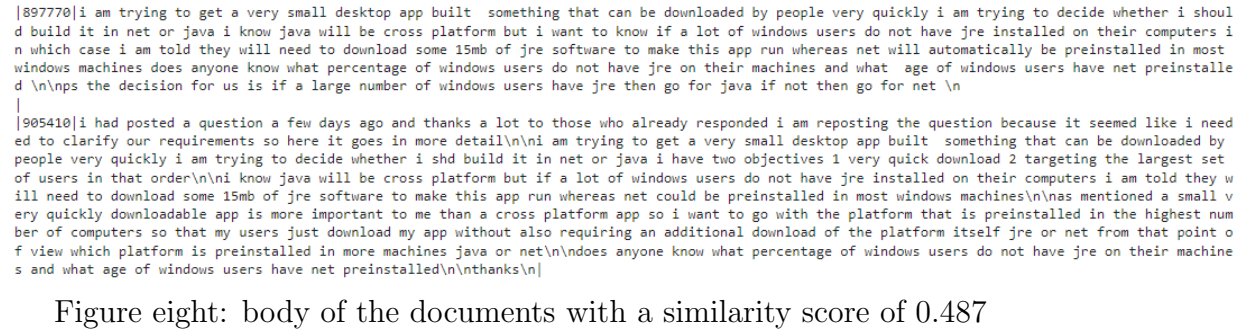
Figure six: body of the pair of the most similar documents

The table in figure six shows the body of the pair of documents having the lowest Jaccard Distance. Then, documents having ids 612820 and 634630 have a similarity score of 0.48, while documents having ids 897770 and 905410 have a similarity score of 0.487.



612820	i have a property grid that helps me manage all of the controls on a form these controls are for designertype folks so im not really worried that much about the user interface until someone selects multiple objects\n\ni have a uitypeeditor for the bottomdiameter property on these common objects it keeps track of units meters vs feet and does some nice things onthefly however when someone selects two or three common objects bottomdiameter is blank even though it evaluates to the same text string\n\nthe reason i think that it is blank is that it is actually three separate objdiameter objects how can i tell the property grid to behave like all of the other properties and show the value if it evaluates to the same string\n\nupdate for example the anchor property has a text output of top right but when you pull it down it is an object yet when you select five objects on your form that all have the same anchor setting you can still see the string top right in the property grid\n
634630	i have a property grid that helps me manage all of the controls on a form these controls are for designertype folks so im not really worried that much about the user interface until someone selects multiple objects\n\ni have a uitypeeditor for the effectivediameter property on these common objects it keeps track of units meters vs feet and does some nice things onthefly however when someone selects two or three common objects effectivediameter is blank even though it evaluates to the same text string\n\nfor example in most controls microsoft has the anchor property that has a text output of top right when you pull it down it is an object with a nice uitypeeditor yet when you select five objects on your form that all have the same anchor setting you can still see the string top right in the property grid\n\ncode\n\nprecode ltsummarygt\n the default containing class for all unitmanagement conversion classes\n ltsummarygt\n\n serializable\n editorattributetypeofumconversiontypeeditor typeofuitypeeditor\n typeconvertertypeofumconversiontypeconverter\n\npublic class umconversion\n\n \n\n\n\npublic class umconversiontypeeditor uitypeeditor\n\n \n\n\n\n\n now in my designer class i have \n private double effectivediameter get set \n\ndisplaynameeffective diameter\npublic virtual umconversion effectivediameter\n\n get\n \n umconversion ret new umconversion effectivediameter\n retmeasureinsi si\n return ret\n \n set\n \n effectivediameter valueimperialunits\n \n \n\ncodepre\n\ncode\n\nif i select several of my custom objects all with the same effective diameter how do i get effectivediameter to display in the propertygrid like anchor does right now that field is always blank\n

Figure seven: body of the documents with a similarity score of 0.48



897770	i am trying to get a very small desktop app built something that can be downloaded by people very quickly i am trying to decide whether i should build it in net or java i know java will be cross platform but i want to know if a lot of windows users do not have jre installed on their computers in which case i am told they will need to download some 15mb of jre software to make this app run whereas net will automatically be preinstalled in most windows machines does anyone know what percentage of windows users do not have jre on their machines and what age of windows users have net preinstalled \n\nps the decision for us is if a large number of windows users have jre then go for java if not then go for net \n
905410	i had posted a question a few days ago and thanks a lot to those who already responded i am reposting the question because it seemed like i needed to clarify our requirements so here it goes in more detail\n\ni am trying to get a very small desktop app built something that can be downloaded by people very quickly i am trying to decide whether i shd build it in net or java i have two objectives 1 very quick download 2 targeting the largest set of users in that order\n\ni know java will be cross platform but if a lot of windows users do not have jre installed on their computers i am told they will need to download some 15mb of jre software to make this app run whereas net could be preinstalled in most windows machines\n\nas mentioned a small very quickly downloadable app is more important to me than a cross platform app so i want to go with the platform that is preinstalled in the highest number of computers so that my users just download my app without also requiring an additional download of the platform itself jre or net from that point of view which platform is preinstalled in more machines java or net\n\ndoes anyone know what percentage of windows users do not have jre on their machines and what age of windows users have net preinstalled\n\nthanks\n

Figure eight: body of the documents with a similarity score of 0.487

By looking at the tables in figures seven and eight, it is possible to visualize the body of the documents having, respectively, a score of 0.48 and 0.487.

6 Conclusions

By looking at the results, it is possible to conclude that two documents could be classified as similar even though the context they belong to is different. In particular, for instance, the bodies of the two documents having the smallest Jaccard distance (0.375) are related to questions about two different things, but they are textually similar. In particular, they have a great amount of words in common, such as 'possible', 'to map', 'using', 'fluent', 'nhibernate'. They also have a comparable style.

7 Personal notes

"I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study."