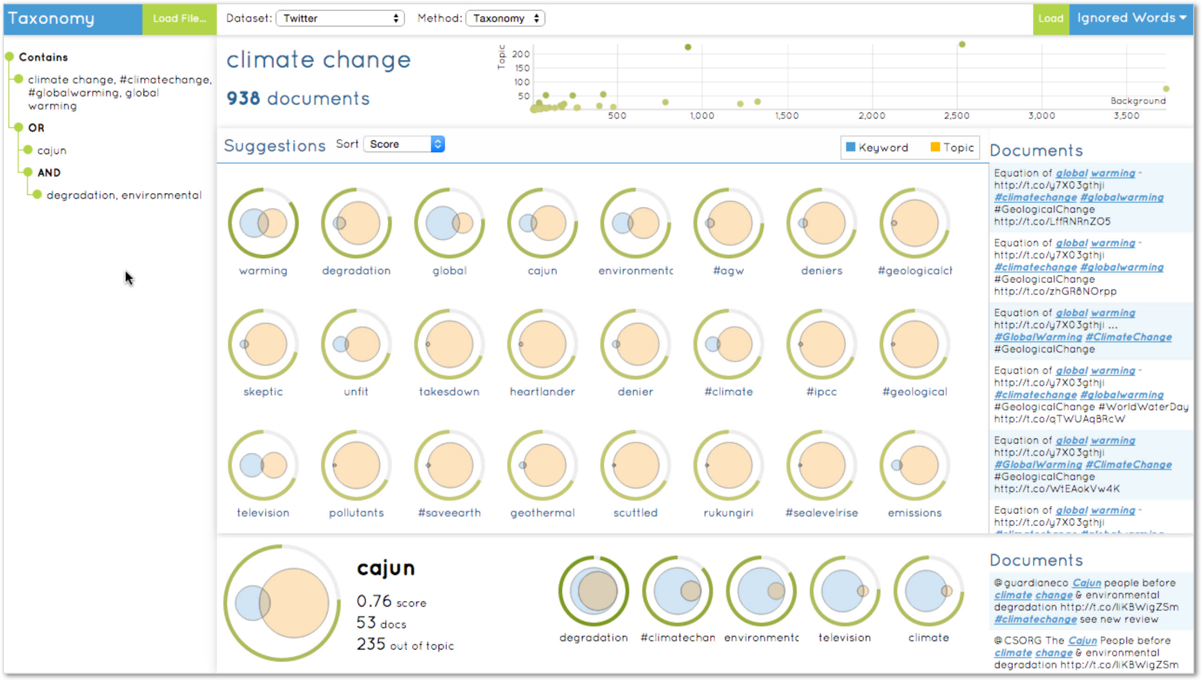
Taxonomy Generator: Generating taxonomies through word suggestions and visualization for the topic climate change

Cristian Felix, Peixin Li, Wei Zhang



Dataset: Twitter

Method: Taxonomy

Contains
- climate change, #climatechange, #globalwarming, global warming
- OR
  - cajun
  - AND
    - degradation, environmental

climate change

938 documents

Sort Score

Equation of global warming - http://t.co/y7X03gthji #climatechange #globalwarming #GeologicalChange http://t.co/LffRNRnZO5

Equation of global warming - http://t.co/y7X03gthji #climatechange #globalwarming #GeologicalChange http://t.co/zhGR8NOrpp

Equation of global warming http://t.co/y7X03gthji ... #GlobalWarming #ClimateChange #GeologicalChange

Equation of global warming - http://t.co/y7X03gthji #climatechange #globalwarming #GeologicalChange #WorldWaterDay http://t.co/qTWUAqBRcW

Equation of global warming http://t.co/y7X03gthji #GlobalWarming #ClimateChange #GeologicalChange http://t.co/WtEAokVw4K

Equation of global warming - http://t.co/y7X03gthji

cajun

0.76 score

53 docs

235 out of topic

@guardianeco Cajun people before climate change & environmental degradation http://t.co/liKBWigZSm #climatechange see new review

@CSORG The Cajun People before climate change & environmental degradation http://t.co/liKBWigZSm

Fig. 1. Taxonomy Generator main screen

Abstract—This project aims to produce an easy and fast way for users to create taxonomies through interaction and visualization. Nowadays many people uses the social media to express their opinion or discuss about many topics. This data could reveal the way how people around this world react to social issues such as environment, pollution, and sea level rising problems. One problem is that it is hard to filter which documents are interesting and which are not. For example, which words belongs to the topic climate change. An easy way would be searching for texts that mention the term climate change, the problem is that some documents may talk about climate change but do not mention the word, for example they could mention warming planet for example. One option is to create a taxonomy for the topic, that is, a description of which words belong to a topic and which ones do not. Come up with this words is a hard task, in this work we propose a tool to help finding this word through corpora exploration based in term suggestion and supported by interactivity visualization, the taxonomy created is then used to perform a query in the documents and get only relevant documents.

Index Terms—Visualization, taxonomy, term suggestion

1 Introduction

Nowadays many people use the social media to express their opinion or discuss about many topics. This data could reveal how people around the world is reacting to social issues such as environment, pollution, and sea level rising problems. The problem is that there are som many documents and it is hard to filter which documents are interesting and which are not. For example, one may be looking for documents that are related to the topic *climate change*, an easy way would be searching for texts that mention the term *"climate change"*, the problem is that some documents may talk about climate change but do not mention these words, for example they could mention *warming planet.*

To solve this problem many approaches have been proposed many of them relying in machine learning algorithms, but the problem with machine learning approaches is that they (1) needs specialized data mining staff (2) do not provide a clear description of what belongs to the topic and what not, not allowing the domain expert to easy interact with it. Another option is to create a taxonomy for the topic, that is, a description of which words belong to a topic and which ones do not. The taxonomy is a description of rules containing boolean expressions and words, and specify combinations of words that have to happen in a document to be considered belonging to the topic. In this way the taxonomy works as a binary classifier returning true if the document

- *Cristian Felix. E-mail: cristian.felix@nyu.edu.*
- *Peixin Li. E-mail: pl1315@nyu.edu*
- *Wei Zhang. E-mail: wz707@nyu.edu*

*Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*

belongs to the topic and false if not.

The problem with taxonomy is how do we create a taxonomy? How do we decide which words are significant to the topic? Take the word *"change"* for instance, it is in the description of the topic *"climate change"* but it is not a discriminative to the topic. Even hard is to think in words that exclude documents of the topic. For instance one may say *"The weather is hot today, I just saw a girl in a short dress"* this is not about climate change.

Our goal in this work is to help people to solve the problem of come up with words or combination of words that belong to a topic, and words that do not belong to it, and while creating the taxonomy, get a sense of if it work or not. To do that we provide an interactive tool based in visualizations where the user can first create a hypothesis of a taxonomy and then through interaction and term suggestion build a more and more complex taxonomy, while seeing results in real time.

In this project we paired up with the UN Global Pulse, a division of United Nations, its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action. They create tools to explore text data and much of the time classify documents using taxonomy.

One example of applicability of the tool is the project "How the World Tweets: Climate Change" [1] where they use a taxonomy to show how people are talking about "climate change" around the word. They created this taxonomy in a system that works like a brainstorm process, where they sat around a table and then suggest possible words, then once this is done, they test the result, and then come back to the taxonomy and changes it which is necessary. One example of how complex the taxonomy can be is available at the same website [2].

2 Related Work

This work touch in three different areas of research, Interactive classifier building, Term or query suggestion and text visualization.

2.1 Interactive classifier and filter creation

This is an area where the goal is not just to train a classifier, but instead build it with constant interaction and feedback from the user. In 1994 Shneiderman [12] proposed dynamic queries, allowing the user to build queries interactively while seeing the intermediate results, our work try to provide the same real time feedback to the user. Scatterblogs2 [1] allows the user to interactively changes parameters in a classifier performing active learning, our work differs from them because in this work the user is not just tuning a trained classifier, but building one. Heimerl et al. [9] also provide an active learning approach.

2.2 Term or query suggestion

Zheng-Jun Zha [18] uses query suggestion to resolve ambiguities of a query, for example differentiate apple the fruit from the computer company, Diane Kelly et al. [10] uses term relevance feedback techniques. Jian Wang and Brian D. Davison[15] provide suggestions through the tags applied to a document. Canini et al.[2] uses tf-idf and LDA to select words to suggest. P. Deepak et al [5] but association rules to discover search engines related queries als has been used [6]. Song, Yang [13] provides suggestions through mining user query reformulation activities. Zhou, Joe[19] uses several methods to do automatic single-word terms, two-words terms and multiple-words terms suggestion. This work can be adapted to different suggestions algorithms but for this implementation we used the Google Normalized Distance[3]

2.3 Text visualization

Viegas FB, Wattenber [14] created a more aesthetic version of tag clouds, a text visualization that is very common used, we avoided tag clouds because although very intuitive, the length of the word affects how the size of the word is perceived, DocBurst [4] provides a semantic exploration of corpora comparing words in a WordNet. TwitInfo [11] uses a common approach to explore tweets presenting a map and some lists containing users, tags and the tweets themselves with sentiment information. HealthMap[7] uses a similar approach to visualize and classify media reports. Wordtree[16] is a tree visualization showing connection between words very useful to display word relations, in this work we show bi-grams in an aligned list, Theme River[8] uses a flow layout to show how topics change along the time. TIARA[17] also presents a topic timeline. Most of this work visualizes documents, we are more focused in visualizing words.

3 Data Analysis and Abstraction

The basic data used is raw text that will be used to build a corpora, this corpora will be used as reference to process and suggest relevant keywords, then the user uses these keywords to construct a taxonomy. In this section we describe these three data itens (text, keywords and taxonomy), their attributes and how they are generated.

3.1 Corpora

The corpora is the most important component of this tool, the quality of keyword suggestions and therefore the taxonomy is totally dependent in the quality of the corpora used during the construction process. We utilized two corporas. The first consists of 8 English million tweets collected during April 2015, we use this dataset because this is one of the main documents being classified by taxonomy at Un global Pulse, however tweets usually has much noise and not always provide good suggestions. For this reason we allow the user to change between different datasets, give that different dataset provide different suggestions, the user has more options of keywords while continue to test the taxonomy in the target dataset. In this work we also provide a list of 7 millions new titles as alternative dataset.

The text is tokenized in two different ways, in the first each word is transformed in a token and indexed, user mentions and hashtags of tweets also are transformed in tokens. To simplify the process URLs are transformed in a special token called _link. The same process is repeated again, but this time each token is composed of 2 words, this bi-gram is called shingle. We use shingles to suggest bi-grams to the user like *"global warming"* and also to explore words that appear just after or before an keyword.

3.2 Keywords

Using the selected corpora the system search and suggest keywords for the specified topic. A keyword is a word or a shingle that the system considered significant to the topic. A keyword is called significant if the distribution of this keyword in documents that satisfy a given search is very different from the distribution of the keyword in the full dataset. For example if the keyword *"global warming"* appears in 5% of the dataset, but in 40% of documents satisfying the search *"climate change"*, we say that *"global warming"* is significant to the search or topic *"climate change"*.

The suggestion algorithm works based in search, in our case this search is the rules of the taxonomy transformed in a query or a search specified by the user. Given a search the suggestion algorithm follows n steps:

1. Select all documents that mach the search criteria, this set is called foreground set while the dataset is called background set.

2. Select all different words in the foreground set

3. For each word in the set compute the number of documents in the foreground that contains this word, and the number of documents that contains this word in the background set,

4. Using the total foreground and background count, together with the background and foreground count for each word, compute the score.

5. Select the top *n* keywords ordered by score in descendant order as suggestions

[1] http://unglobalpulse.net/climate/

[2] http://unglobalpulse.net/climate/taxonomy/

As result from the algorithm we have a table where each row correspond to an suggested keyword with 3 attributes: Score, foreground count and background count. Also we have the total foreground count and background count, 1 summarize these attributes. To allow the interactivity the keyword suggestion has to be able to provide suggestion in a short amount of time.

As score algorithm we used the Google Normalized Distance (GND)[3], it provides some benefits. Firs it is able to suggest some rare terms, this is good because some words that describe the topic may have been used just few times. This has the drawback of also suggesting misspellings, but given that the user can decide if he or she wants to use the keyword, we do not see that as a big problem. Also this algorithm provides a number between 0 and 1, this helps to keep the visualization for the same keyword consistent while the search changes over the time.

The GND algorithm is given by:

$$GND(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logM - min\{logf(x), logf(y)\}}$$

Where $f(x)$ is the number of documents in the foreground set, $f(y)$ is the number of documents in the background set, $f(x,y)$ is the intersection between then or the number of documents in the foreground set that contains the keyword, and M is the number of documents in the corpora. The GND algorithm score relevant terms low, therefore we have to invert the order of terms by performing:

$$Score = e^{-GND(x,y)}$$

This will result in a number from 0 to 1, where 1 means very significant and 0 no significance.

3.3 Taxonomy

Using the keyword suggested by the system together with other that he or she came up during the exploration process, the user builds the taxonomy. The Taxonomy consists on a tree containing keywords and the boolean operators *AND*, *OR* and *NOT*. Each node of the tree is a rule of one type of operation, containing a list of words and a list of other rules. The list of words are implicit OR operation, this means that the list of words return true if any of the words exists in the document.

To classify a document a document the tree is traversed where the Boolean result of each children is propagated back to the parent that compares with the list of words it has and propagates up again. If at the end of the process the root node propagate **true** the document is classified as satisfying the taxonomy.

The taxonomy can be converted to different types of SQL queries but also no SQL queries or even Regex expressions.

4 Task Analysis and Questions

In our meeting with UN global Pule, one aspect of this project was clear, the target user not necessarily will be a computer science expert, therefore the tool has to be intuitive, easy to use, and allow the user to tune the taxonomy without having to relay in mathematics or many parameters. Based in this idea, we aim in developing a tool that helps the user answer the following questions in a easy way.

1. Given a topic, what combinations of words are significant for this topic?
   This is the core of the tool, the hardest part of the process of taxonomy creation is to come up with words, here the system should suggest words that the user did not though about.

2. Given my current taxonomy, there are documents that are being matched but should not?
   Many times the user create a good taxonomy, but the dataset has to much noise, and it is important to the user has to be able to find these noisy and fix the taxonomy.

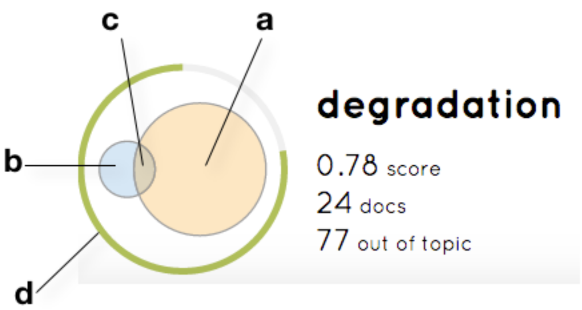


Fig. 2. Venn Diagram Glyph for the word degradation in the topic climate change - (a) Number of documents in the scope set(topic), (b) Number of documents that contains the selected word, (c) number of documents that in the topic that also contains the keyword, (d) score of the keyword for the topic

3. Given a word and a topic, why this word was suggested? This is important, because to decide if a word should or not be part of the taxonomy it is important to know the relevance, the score is the main indicator of relevance, but the user has to be able to understand why a word got an specific score.

4. Given a word and a topic, how this word is used inside the topic and which other keywords are related to this word? This is an important task, because allowing the user to explore the word helps him to make sense of the word, and also create new keywords from the current keyword.

5 Visualization and Interaction Design

The design process of the tool was developed in an interactive way, incorporating feedback provided by user to whom we presented the tool, through out this process we were able to learn what was not working, and also come up with ideas that could help to improve the tool. In This section we present the tool and the its possible interactions.

5.1 User Interface and Visualization

The main interface of the tool is presented in the **??**, the tool is divided in 2 main views, the taxonomy and the exploration view. Through the exploration view the user can come up with keywords that he or she moves to the Taxonomy View where the taxonomy is being built.

5.1.1 Venn Diagram Glyph

The venn diagram glyph is the basic component of the interface of the tool it is a venn diagram surrounded by an arc. Figure 2 shows the glyph in details for the word degradation.

We use a venn diagram to represent the relations between 3 variables: background count, topic count and intersection count, the main reason to choose a venn diagram is the familiarity, but also it performs very well when comparing 2 sets. However it not work to compare more than 2 sets. In this diagram one circle represents the keyword and the other circle represent the scope that we want to compare this keyword. In most of the cases the scope is the documents that satisfy the taxonomy, but this is not always true, we also use it to compare two words, in this case the circle that represents the scope, will be the documents containing the word being compared.

The set of documents containing the keyword is represented by an blue circle, while the set of documents satisfying the scope is represented by an orange circle, the area of each circle is mapped to the number of documents in each set. both circles are aligned ate the y-axis, and the distance between the centers of the circle is calculated such that the area created by the overlay of both is equal to the number of documents present in both sets, The overlay area is colored with a mixture of the colors of both circles.

Table 1. Keyword attributes

Attribute	Type	Range	Description
keyword	Categorical	As many different word are in the corpora	It is the label identifying the keyword
Foreground Count	Numeric	Goes from 0 to the #documents in the corpora	It is the #documents in the foreground set containing the keyword
Background Count	Numeric	Goes from 0 to #documents in the corpora	#documents in the background set/corpora containing the keyword
Score	Numeric	Changes depending on the algorithm, 0 to 1 for GND	Represents how significant the key word is for the search

The venn diagram uses area, so it works well comparing sets with very different sizes. First because the area distributes the count in a 2D space while using lines or bars will distribute the length just in 1D. Also a very small set may be just a point but still being perceived in contrast with others versions we tried where a small set becomes a tiny line and sometimes the user can't see it.

The glyph helps to easily make some comparisons, for example one can easily see the size of the two sets, we can detected for example if a set of documents containing one keyword is to small compared with the topic, adding this word may actually making the taxonomy to overfit the dataset. Another example, in some cases one circle is totally inside another, if the circle representing the word is totally inside the topic, this means that at the moment this word is redundant to the topic, and adding it will not increase the number of documents being matched by the taxonomy.

Another benefit of this representation is that it use a scale that depends only on the topic and the current keyword, we tried different visualizations where we use a common scale along all words being visualized, but even using logarithmic scale, the difference of the size of too many sets resulted in some words affecting the visualization just because they are too frequent. Also the size of the circles keep more or less steady along the exploration, and it is not affected if words are added or removed from the list of suggested words, this is important, because in this case the sizes only changes if the topic set changes.

One drawback of this approach is that the size of the circle representing the topic, changes from one keyword to another, even if the number of documents is the same. But we believe that the user do not lose the sense of how frequent a word is, this happens because every word can be compared with the topic, and the difference from the topic tells us the frequency of the word. From the figure above is easy to say that the frequency of **climate** is higher then **#climatechange** and this has a higher frequency than **degradation**. The main goal of this representation is to provide comparison between the two sets, ant for this tasks it works very well, therefore we choose to use this approach even with the problem described.

The venn diagram is surrounded by an arc, the angle of the arc is mapped to the score that the keyword received, to facilitate the exploration, we also mapped the color of the arc to the score, making much easier to detect word with very high score and word with very low score and also get an overall idea of the score distribution when you have many glyph side by side.

5.1.2 Taxonomy View

The taxonomy view is an representation of the taxonomy being built, the visualization is based in a tree, following the same patter of folders trees in file manager systems (Windows Explorer on Windows or Finder on MacOS). Each node is positioned from top to bottom and it is indented according with its level on the tree, a green circle represents a node, a line is drown connecting the child node with its parent. The words describing rules(boolean operators) are in bold.

We selected this options for some reasons. First the user is familiar with this structure since it is the same present in her operational system, second it is easy to read, the user can read from top to bottom and easily spot the rules and which words belongs to each rule, this representation also grows mostly in a vertical way, even in trees with 4 or 5 levels, it still narrow, in this way it will be very rare cases when the user will need a horizontal scroll bar, having only to deal with the vertical one, this also makes easy the drag and drop process of adding words, this process will be detailed in the interaction section.

5.1.3 Exploration View

The exploration view is divided in 3 parts, the top one presents a summary stating the name of the topic being created and how many documents the current taxonomy matches in the selected dataset. The middle contains the results of the suggestion algorithm, that is, the suggested keywords, and a sample of documents satisfying the taxonomy. The bottom shows details of the current selected word. In an intuitive way, the level of granularity grows from top to bottom. Result overview → keyword overview → keyword detail.

The scatter plot in the top view shows all suggested keywords as dots with the x-axis mapped to the number of documents that contains the word in the database, the y-axis is mapped to the number of documents containing the word that **satisfy** the taxonomy. Recall that these are the two keyword specific variables used to compute the score, the other two are the same for all key words (Document Total, Taxonomy Total), for this reason the position of the key word in the scatter plot will be strongly connected with the score. The score of the word is mapped to the color in a divergent scale going from orange to yellow (0 to 0.5) to green (0.5 to 1), we choose this scale because it makes easy to the user spot words with score below 0.5.

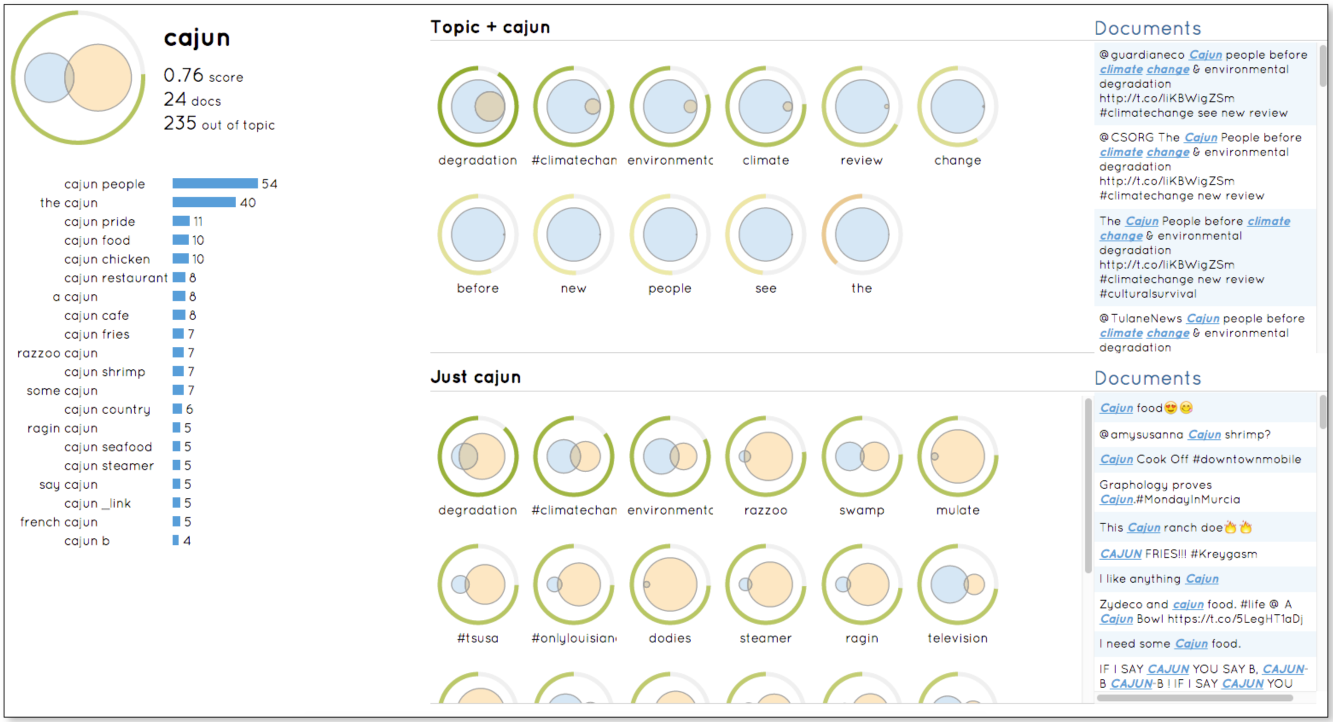
The middle of the the exploration view shows the suggested keywords represented by the venn diagram glyph, the keywords are ordered by the score in a fluid layout, from left to right and them top to bottom. In the glyph the keyword is mapped to the blue circle and the topic to the orange circle. This section also present a document view with the top most relevant documents for the taxonomy, the words that were used to selected the document are highlighted.

When a keyword is selected the bottom part shows details of this word. It presents the keyword glyph together with the actual numbers for the score, foreground count and background count. Also some words that are discriminative for the selected word in the topic are presented with their respective glyph. A list of documents matching the taxonomy and containing the word are shown in the left of this section, with the relevant words highlighted.

5.1.4 Keyword details view

The user can open the Keyword details view by clicking in more in the bottom of the screen. This will expand the current details of the word showing additional information. The Figure 3 shows this view for the word cajun. This view presents the gliphy of the word together with a **bi-gram bar** chart, this bar chart shows bi-grams that contains the selected word, and each bar is mapped to the number of documents that contains the bi-gram in the dataset. The bi-grams are centered in the selected word, making easy to see if the other word in the bi-gram comes after or before the selected word.

This view also shows a two different sections containing keywords with glyph and documents list. The top one shows suggested words that are related with both, the current taxonomy and the selected word, we use this view to get a sense of *how* and *why* the word is related with the topic. The second one shows the suggested keywords that are related with just the word, not taking in consideration the topic, this is also important because it shows how the word is used in general, helping to get a better sense of the keyword.



cajun people	54
the cajun	40
cajun pride	11
cajun food	10
cajun chicken	10
cajun restaurant	8
a cajun	8
cajun cafe	8
cajun fries	7
razzoo cajun	7
cajun shrimp	7
some cajun	7
cajun country	6
ragin cajun	5
cajun seafood	5
cajun steamer	5
say cajun	5
cajun _link	5
french cajun	5
cajun b	4

Documents

@guardianeco Cajun people before climate change & environmental degradation http://t.co/liKBWigZSm #climatechange see new review

@CSORG The Cajun People before climate change & environmental degradation http://t.co/liKBWigZSm #climatechange new review

The Cajun People before climate change & environmental degradation http://t.co/liKBWigZSm #climatechange new review #culturalsurvival

@TulaneNews Cajun people before climate change & environmental degradation

Documents

Cajun food😍😋

@amysusanna Cajun shrimp?

Cajun Cook Off #downtownmobile

Graphology proves Cajun.#MondayInMurcia

This Cajun ranch doe🔥🔥

CAJUN FRIES!!! #Kreygasm

I like anything Cajun

Zydeco and cajun food. #life @ A Cajun Bowl https://t.co/5LegHT1aDj

I need some Cajun food.

IF I SAY CAJUN YOU SAY B, CAJUN-B CAJUN-B ! IF I SAY CAJUN YOU

Fig. 3. Keyword detail for the word cajun in the topic climate change

5.2 Interaction

The interaction plays a very important role in this tool, the basic pipeline is (1) Provide a search criteria to the system, (2) the system returns a list of suggested keywords, (3) The user explore keywords, (4)The relevant keywords are added to the taxonomy. This process keep repeating until the user is satisfied with the taxonomy resultant. In this section we present details of how the user interact with the tool to perform this 4 stages.

5.2.1 Search

The search is the process of generating a subset of documents that will be the scope of the suggested keywords. To perform a search the user can use the load button. By default the search is performed converting the taxonomy to a query and ruining it against the current selected dataset in the dataset combobox.

The user can change the dataset at any time, since the taxonomy is not connected with any dataset. Words in the taxonomy can be deactivated, this means that this words will not be used to generate the search, although it still part of the taxonomy, this is done by clicking on the word that the user wants to deactivate. Deactivated keywords are shown in light gray, to reactivate the word the user can click on it again.

The user can change the method to search, in this case a input box will be shown where the user can perform arbitrary searches, the searches typed in this box do not affect the taxonomy, and when using this method the taxonomy will not be considered while performing the search.

5.3 Exploring Keywords

The system will plot the suggested words using the venn diagram glyph, a scatter plot and lists of documents in the view as described in the previous section.

It is possible to highlight the words in the scatter plot by moving the mouse over the desired dot, the word win the list of suggested words will be highlighted. It is also possible to select a word by clicking on it.

In the keyword list the user can also click to select the word, details of the selected word will appear on the bottom of screen. words in this list can be reordered through the sort combobox, the options are by score, number of documents in the topic, and number of documents overall.

The user can also ignore words in the list, telling the system that this word is not relevant, this can be done by dragging the keyword and drooping it on the *Ignored words* blue rectangle, if the user change his mind he can bring the word back by hovering the retagle and clicking in the trash icon aside the keyword that he or she wants to bring back.

When a keyword is selected the user can open a new view that shows more details of the selected word by clicking in the more button on the bottom of screen.

5.4 Editing the Taxonomy

The user can change the taxonomy by adding rules and words to the taxonomy and removing then.

To add a word to the taxonomy the user can drag and drop the desired keyword to the node in the taxonomy that he or she wants this word. The user can also add words that were not suggested by hovering the desired node and clicking in the link add, the system will ask the user to type the desired word. To remove a word the user can double click the word and the system will remove it.

To add a rule, the user can over the node and select the desired boolean operation (AND, OR, NOT), a new node will be added to the taxonomy where the user can start to add new words. To remove a rule, the user can hover the rule and click on the link remove. The rule will be removed with all words and children it has.

6 Findings and Insights

For the current prototype used 8 million tweets collected by us during the month of April. The system worked very well providing suggestion in a good timing and relevant to the topic, we also uploaded the dataset provided by Accern that contains titles of articles of news, this helped us to see how changing the dataset changes the suggestions.

6.0.1 Exploring climate change

As a simple test we decided to explore how to use the tool to create a taxonomy to climate change. After typing the topic we get a list of 50 suggested keywords. By the scatter plot was easy to spot some words that were outliers, these words turned out to be *climate* and *change* the two words that define the topic, this helps because the system recognizes ”climate change” as one keyword, and because it do not know if we want to add *climate* and *change* with other words, it shows this words until either we add then separated in the taxonomy or ignore them.

The list also contained hashtags like #climatechange, #savetheearth, and #globalwarming all clearly related with the topic. Also words like *warming, carbon, greenpeace, rainforest* However it also suggested two words that we do not even know what they were.

The first were “*cajun*”, by looking to the glyph we saw that around one third of the time this word were used, it was in the climate change topic, by clicking on it we got a list of related keywords containing for example degradation and environmental. Looking in documents in the topic that contains this word, we found a link that redirected us to an article talking about the Cajun tribe that is loosing land because of the rise of water. Before adding this word to the taxonomy we looked in the bi-gram chart and discovered that people often talk also of their food, therefore if adding this word to the taxonomy, would be important also add constrains that checks if the document is not about their food.

The second was “*rukungiri*”, by looking in the glyph we saw that it was a very rare word, because it circle was too small compared with the topic, whe we selected the word, we saw that the most significant word for it was forestry, we then opened the details of the word, and from the bi-gram chart we saw that almost always it was followed by the word “*district*” looking in the text in the document view, we discovered that it was a district commemorating the forest day. We discovered that although the word get a high score, these do not means that it should go in the taxonomy, the high score was just because they were celebrating a holy day in the exact period we collected the data, and people from there tweeted only about the holiday.

7 Limitations and Future Work

7.1 Limitations

Although the tool already stated to show some good results it still suffering from some limitations. First the tool is very biased towards the dataset, in our case we used twitter data and these social network allows people to retweet information, therefore duplicating documents. The suggestion algorithm do not take this int account, therefore giving higher score to words in these repeated documents.

Also the twitter dataset has so many noise, and even with more than 8 million tweets, we were able to get only 431 documents containing the term climate change. Much of the tweets were repeated or did not contained relevant information, Even with the data we got from Accern we got 1487 documents containing this word of more than 7 million documents.

This means that for this tool to be generic for all kinds of topics we need a very huge number of documents, this increase the time that the algorithm takes to suggesting words, therefore limiting the interactivity. On option is to use a higher number of different datasets, perhaps from news.

Another limitations is that right now, it is not possible to import and export the taxonomies, limiting the use of the tool. Also the language of the documents is a limitation, we did not test with other foreign languages. This is an important feature for UN Global Pulse due to their global actuation.

Our tool do not take in consideration the time and space of the documents, for example if someone want to mach documents about some specific event would be good to use filter the dataset to documents published in the same period and geography of the event.

7.2 Future Work

There is plenty of improvements that can be done in this tool. For examples exploring different suggestions algorithms trying to understand which one perform best for each task, and how they perform and datasets with large volumes of documents.

The language is also a place to be explored, we want to understand if this tool can be used to create taxonomies in different languages.

We believe that time also plays a important role, explore how to show the changes of the relation between the word and the topic helps to understand if the word is connected with the topic in a long term like *global warming* in the topic *climate change* or temporary, like *rukungiri* and *cajun* to the same topic.

8 Conclusion and Lessons Learned

Extracting and visualizing text is not an easy task, and this becomes even harder when the processing has to be online, the tool presented in this work strongly relays in interaction and online text processing. The project helped us to understand the problem and touch in a research problem that we believe is very relevant.

From the visualization design process point of view we had o define visualization problem, we have to provide a convincing problem statement. We should answer the following question to make the problem statement convincing. What is the problem you want to solve? Why is it important? Who has this problem? Do you have direct access to end-users? How do your end-users currently solve this problem? What impact may your system have on the target users? It is very important to show that you have researched the problem thoroughly by interacting with your end-users and/or by having researched the problem in detail yourself.

For visualization sketch, we choose the map visualization at beginning, and we find there’s no clear justification for why a map is needed. Geolocation does not seem to be a crucial parameter for the problem we want to solve. Then we changed to different glyph and designs until select the venn which is more direct way to show the relation between word.

To add the interaction between user and visualization interface. We create taxonomy generator part, which allow user to add a and, or, not, ignore word, and allow user to create complex taxonomy by add and, or, not to particular word. What’s more, we show the detail about each suggested word, and related twitter.

We believe in the importance of this work, and that there are many research problems opened in both fields, text processing and text visualization, but we think that this work contributed to this fields with some approaches solutions for these problems

9 Link

In this section, we provide the link to github page, working demo and video demo:

1. Github: https://github.com/NYU-CS6313-Projects/Taxonomy-Generator.
2. Working demo: http://nyu-cs6313-projects.github.io/Taxonomy-Generator/index.html.
3. Video demo: https://vimeo.com/127993895.

References

[1] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2022–2031, 2013.

[2] K. R. Canini, B. Suh, and P. L. Pirolli. Finding credible information sources in social networks based on content and social structure. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 1–8. IEEE, 2011.

[3] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

[4] C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. In *Computer graphics forum*, volume 28, pages 1039–1046. Wiley Online Library, 2009.

[5] P. Deepak, S. Chakraborti, and D. Khemani. Query suggestions for textual problem solution repositories. In *Advances in Information Retrieval*, pages 569–581. Springer, 2013.

[6] B. M. Fonseca, P. B. Golgher, E. S. de Moura, and N. Ziviani. Using association rules to discover search engines related queries. In *Web Congress, 2003. Proceedings. First Latin American*, pages 66–71. IEEE, 2003.

[7] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, 2008.

[8] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[9] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2839–2848, 2012.

[10] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 371–378. ACM, 2009.

[11] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM, 2011.

[12] B. Shneiderman. Dynamic queries for visual information seeking. *Software, IEEE*, 11(6):70–77, 1994.

[13] Y. Song, D. Zhou, and L.-w. He. Query suggestion by constructing term-transition graphs. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 353–362. ACM, 2012.

[14] F. B. Viégas and M. Wattenberg. Timelines tag clouds and the case for vernacular visualization. *interactions*, 15(4):49–52, 2008.

[15] J. Wang and B. D. Davison. Explorations in tag suggestion and query expansion. In *Proceedings of the 2008 ACM workshop on Search in social media*, pages 43–50. ACM, 2008.

[16] M. Wattenberg and F. B. Viégas. The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1221–1228, 2008.

[17] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162. ACM, 2010.

[18] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 15–24. ACM, 2009.

[19] J. Zhou and P. Dapkus. Automatic suggestion of significant terms for a predefined topic. In *IN PROCEEDINGS OF THE 3RD WORKSHOP ON VERY LARGE CORPORA, ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*. Citeseer, 1995.