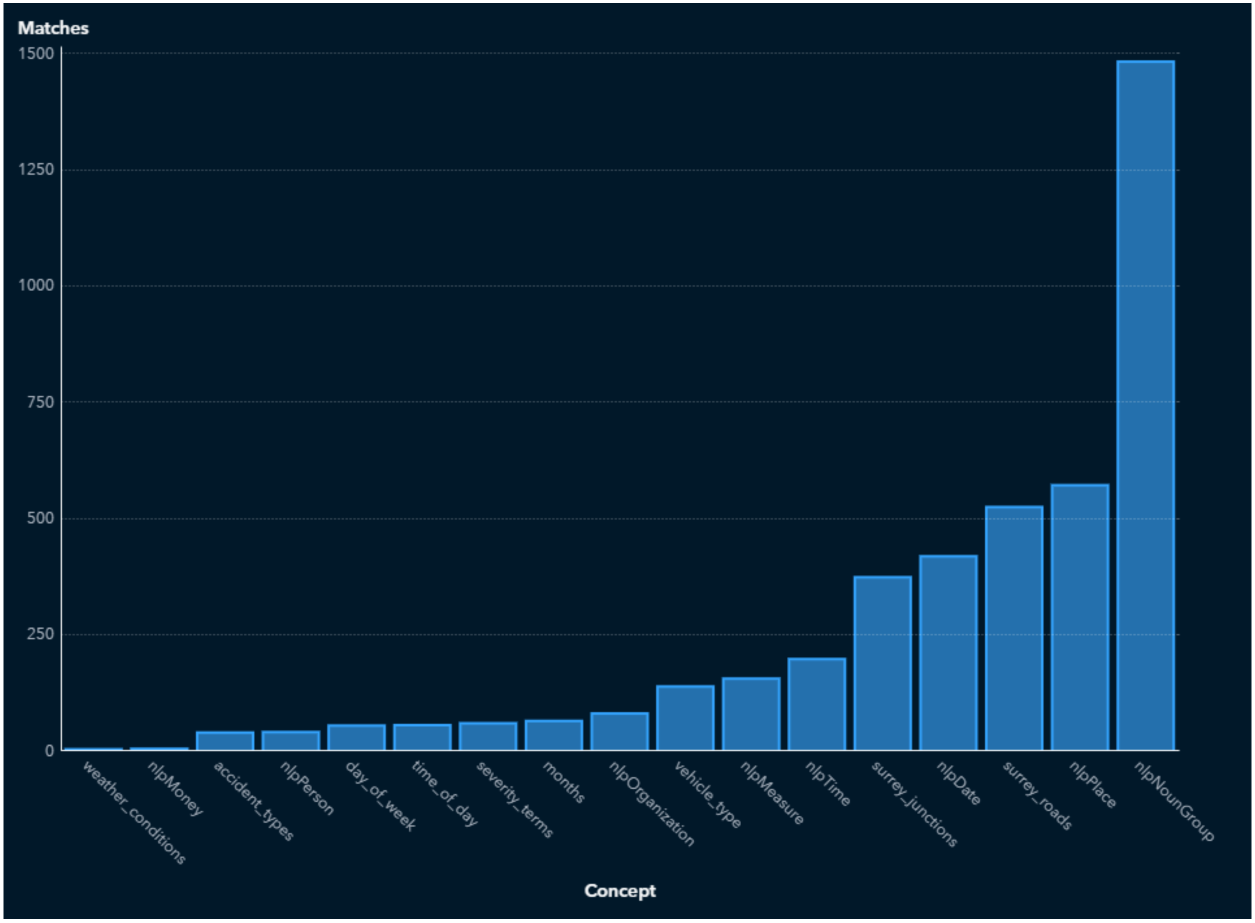
Surrey_Accident_Analysis_Final

"Concepts" Results

by: di00222@surrey.ac.uk

Contents

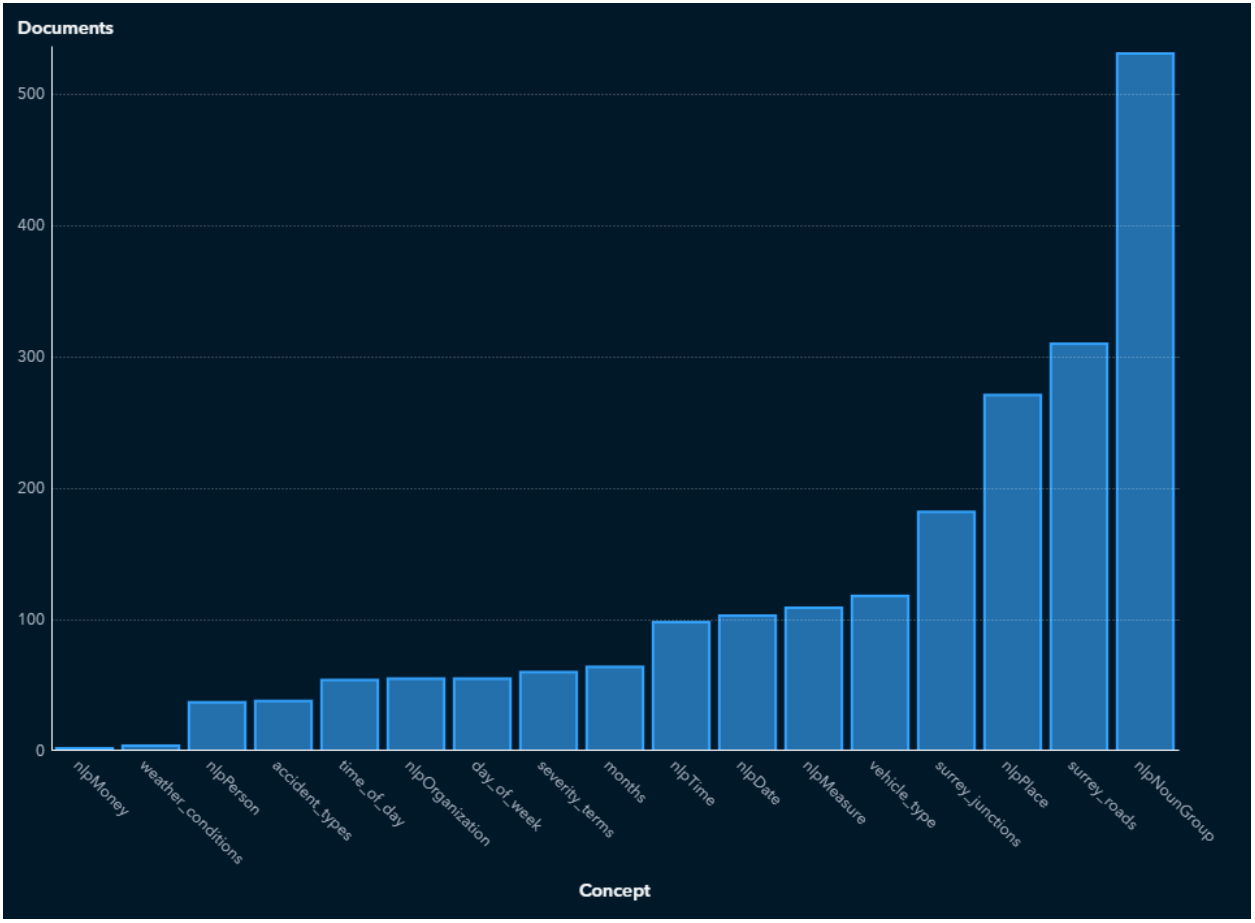
Number of Matches Per Concept



The Number of Matches per Concept report depicts how useful each concept is for finding information in the data generally. The top matching concept in this data, nlpNounGroup, has 1,483 matches, while the least matched concept, weather_conditions, has 4 matches.

This information indicates how closely each concept, and the data in the documents are aligned. Many matches show that a concept is well-defined to extract information from the data set. Fewer matches for a concept indicate that either the data is not appropriate for the concept or that the concept needs further definition by adding or refining rules.

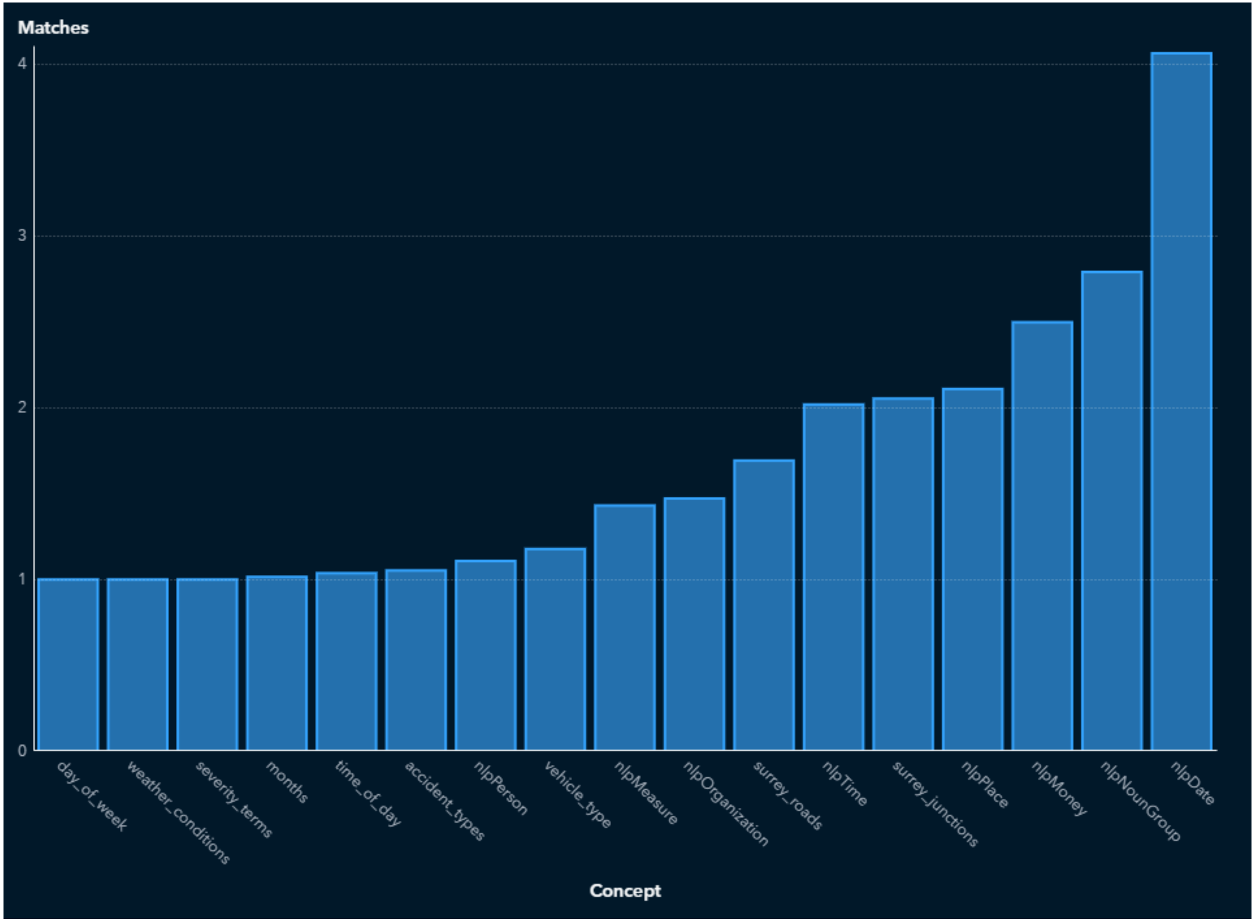
Number of Documents Per Concept



The Number of Documents per Concept report depicts how well each concept covers the documents in the data set. In this data set, the two concepts with the greatest coverage are nlpNounGroup with matches spanning across 531 documents (88.8% of the total documents), and surrey_roads with matches spanning across 310 documents (51.84% of total). The concepts with very light coverage (less than 5%) across documents in this data set are: weather_conditions, nlpMoney.

This information indicates how broad each concept is in terms of how many documents it matches. In a project that is expected to cover all the documents with each concept or a subset of concepts, this report can be used to gauge how close the model is to that goal and which concepts are more complete in their coverage.

Average Number of Matches Per Document



The Average Number of Matches per Document report depicts the amount of information extraction performed by the concept inside each document where it matches. In this data set, for example, when nlpNounGroup matches in a document, it matches 2.79 times on average.

The calculation of the average does not include documents where the match count is zero for that concept. This report shows when a concept is getting the desired depth to help with prioritization during development of concepts within the model.

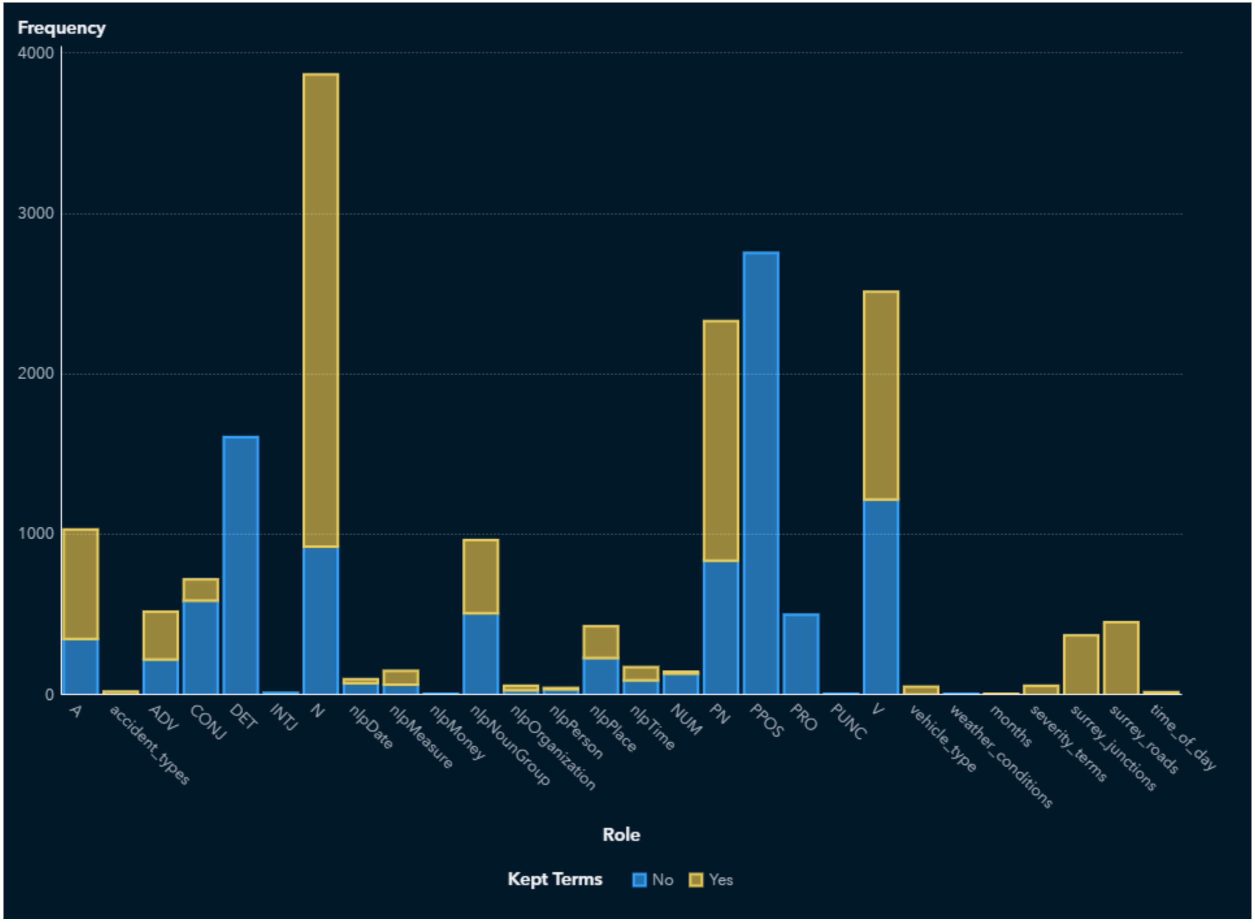
Surrey_Accident_Analysis_Final

"Text Parsing" Results

by: di00222@surrey.ac.uk

Contents

Role by Frequency



The Role by Frequency report is a visual summary of the terms data. Parsing assigns role labels to each term in the entire data set. The height of the bars in this report indicates the frequency for each type of role. In this data set, the most frequently occurring roles are N (noun) and PPOS (preposition/postposition). Together they represent 35.02% of the terms. Note that both bare numbers and most types of punctuation (except some symbols) are removed by default from the terms list and are not reported here. To work with the table of values further, download the data.

For each bar, the areas defined by the two colors represent the proportion of kept or dropped terms with the given role. A downstream Topics node will use only the kept terms to determine topics across the data set. In this data set, 53.94% of the terms that are found in the data have been dropped from consideration in building topics. The dropped terms were removed because of their presence in a stop list or because they do not meet the frequency threshold set by the ‘minimum number of documents’ parameter. Additional terms can be dropped in the interactive view or by adding them to the applied stop list.

Parsing identifies terms through a sequence of NLP (natural language processing) steps, including tokenization, multiword identification, lemmatization, part-of-speech tagging, and noun group extraction. Synonym lists and misspelling detection can be added to the analysis to further group term variants under a single parent term. Concept extraction may also be applied using a preceding Concepts node.

Each resulting role is either a part-of-speech (content or function word) or a concept. Parts-of-speech include labels such as N (noun), CONJ (conjunction), PN (proper noun), ADV (adverb), NUM (numeric), PUNC (punctuation), and so on. Concepts may include noun groups (nlpNounGroup) and those defined and passed forward from a preceding Concepts node in the pipeline.

Descriptive Statistics

Measure	Terms in a Sentence	Terms in a Document
Minimum	1	1
Maximum	52	66
Mean	14.5716	31.6288

The Descriptive Statistics table records patterns that are found across the data set. The average length of sentences by count of terms in this data set is 14.57. The range of sentence length is 1 - 52 terms. The average length of documents by a count of terms in this data set is 31.63. The range of document length is 1 - 66 terms.

The information in this chart can identify unexpected data characteristics that may need to be investigated. The information presented in the table is a summary view only. For more detailed information or to compare data sets, run the profileText action.

Surrey_Accident_Analysis_Final

"Sentiment" Results

by: di00222@surrey.ac.uk

Contents

Sentiment Score Code

```
/****************************************************************
* SAS Visual Text Analytics
* Sentiment Score Code
*
* Modify the following macro variables to match your needs.
**************************************************************/

/* specifies CAS library information for the CAS table that you would like to score. You
must modify the value to provide the name of the library that contains the table to be
scored. */
%let input_caslib_name = "{input_caslib_name}";

/* specifies the CAS table you would like to score. You must modify the value to provide
the name of the input table, such as "MyTable". Do not include an extension. */
%let input_table_name = "{input_cas_table_name}";

/* specifies the column in the CAS table that contains a unique document identifier. You
must modify the value to provide the name of the document identifer column in the
table. */
%let key_column = "{doc_id_column_name}";

/* specifies the column in the CAS table that contains the text data to score. You must
modify the value to provide the name of the text column in the table. */
%let document_column = "{text_column_name}";

/* specifies the CAS library to write the score output tables. You must modify the value
to provide the name of the library that will contain the output tables that the score code
produces. */
%let output_caslib_name = "{output_caslib_name}";

/* specifies the sentiment output CAS table to produce */
%let output_sentiment_table_name = "out_sentiment";

/* specifies the matches output CAS table to produce */
%let output_matches_table_name = "out_sent_matches";

/* specifies the features output CAS table to produce */
```

```
%let output_features_table_name = "out_sent_features";

/* specifies the language of the associated SAS Visual Text Analytics project. This
should be set automatically to the language you selected when you created your
project. */
%let language = "ENGLISH";

/* specifies the hostname for the CAS server. This should be set automatically to the
host for the associated SAS Visual Text Analytics project. */
%let cas_server_hostname = "sas-cas-server-default-client";

/* specifies the port for the CAS server. This should be set automatically to the host for
the associated SAS Visual Text Analytics project. */
%let cas_server_port = 5570;

/* creates a session */
cas sascas1 host=&cas_server_hostname port=&cas_server_port;
libname sascas1 cas sessref=sascas1 datalimit=all;

/* calls the scoring action */
proc cas;
session sascas1;
loadactionset "sentimentAnalysis";

action applySent;
param
table={caslib=&input_caslib_name, name=&input_table_name}
docId=&key_column
text=&document_column
language=&language
casOut={caslib=&output_caslib_name, name=&output_sentiment_table_name,
replace=TRUE}
matchOut={caslib=&output_caslib_name, name=&output_matches_table_name,
replace=TRUE}
featureOut={caslib=&output_caslib_name, name=&output_features_table_name,
replace=TRUE}
;
run;
quit;
```

Execution Environment

Author:	di00222@surrey.ac.uk
File:	File: /export/viya/homes/di00222@surrey.ac.uk/casuser/Sentiment Analysis.sas
SAS Context:	SAS Studio compute context
SAS Version:	V.04.00M0P091624
SAS Client:	SAS® Studio 6.0
SAS Locale:	
Submission Time:	May 11, 2025, 5:13:15 AM
Time Zone:	GMT+01:00
User Agent:	Chrome 135.0.0.0

Code:Sentiment Analysis.sas

```
/*****************************************************************
* SAS Visual Text Analytics
* Sentiment Score Code
*
* Modify the following macro variables to match your needs.
****************************************************************/

/* specifies CAS library information for the CAS table that you would like to score. You must modify the value to provide the name of the l
%let input_caslib_name = "CASUSER";

/* specifies the CAS table you would like to score. You must modify the value to provide the name of the input table, such as "MyTable". Do
%let input_table_name = "TWEETS";

/* specifies the column in the CAS table that contains a unique document identifier. You must modify the value to provide the name of the d
%let key_column = "ID";

/* specifies the column in the CAS table that contains the text data to score. You must modify the value to provide the name of the text co
%let document_column = "Text";

/* specifies the CAS library to write the score output tables. You must modify the value to provide the name of the library that will conta
%let output_caslib_name = "CASUSER";

/* specifies the sentiment output CAS table to produce */
%let output_sentiment_table_name = "out_sentiment_1";

/* specifies the matches output CAS table to produce */
%let output_matches_table_name = "out_sent_matches_2";

/* specifies the features output CAS table to produce */
%let output_features_table_name = "out_sent_features_3";

/* specifies the language of the associated SAS Visual Text Analytics project. This should be set automatically to the language you selecte
%let language = "ENGLISH";

/* specifies the hostname for the CAS server. This should be set automatically to the host for the associated SAS Visual Text Analytics pro
%let cas_server_hostname = "sas-cas-server-default-client";

/* specifies the port for the CAS server. This should be set automatically to the host for the associated SAS Visual Text Analytics project
%let cas_server_port = 5570;

/* creates a session */
cas sascas1 host=&cas_server_hostname port=&cas_server_port;
libname sascas1 cas sessref=sascas1 datalimit=all;

/* calls the scoring action */
proc cas;
    session sascas1;
    loadactionset "sentimentAnalysis";

    action applySent;
        param
            table={caslib=&input_caslib_name, name=&input_table_name}
            docId=&key_column
            text=&document_column
            language=&language
            casOut={caslib=&output_caslib_name, name=&output_sentiment_table_name, replace=TRUE}
            matchOut={caslib=&output_caslib_name, name=&output_matches_table_name, replace=TRUE}
            featureOut={caslib=&output_caslib_name, name=&output_features_table_name, replace=TRUE}
        ;
    run;
quit;
```

Log:Sentiment Analysis.sas

```
1    /* region: Generated preamble */
2    /* Make sure the current directory is writable */
3    data _null_;
4        length rc 4;
5        %let tworkloc="%sysfunc(getoption(work))";
6        rc=dlgcdir(&tworkloc);
7    run;
NOTE: The current working directory is now
      "/opt/sas/viya/config/var/tmp/compsrv/default/2eb57acd-96d9-4ec7-ac12-6a11ff1ae086/SAS_work155B00000206_sas-compute-server-303
      840af-cd9c-4da5-ba07-1e4c1a765b6f-20257".
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
      

8     
9     /* Setup options */
10    title;
11    footnote;
12    options validvarname=any;
13    options validmemname=extend;
14    options dtreset date number;
15    options device=png;
16    
17    /* Setup macro variables */
18    %let syscc=0;
19    %let _clientapp = %nrquote(%nrstr(SAS Studio));
20    %let _clientappabbrev = %nrquote(%nrstr(Studio));
21    %let _clientappversion=2024.09;
22    %let _clientversion=;
23    %let _sasservername=&SYSHOSTNAME;
24    %let _sashostname=&SYSHOSTNAME;
25    %let _sasprogramfilehost=&SYSHOSTNAME;
26    %let _clientuserid = %nrquote(%nrstr(di00222@surrey.ac.uk));
27    %let _clientusername = %nrquote(%nrstr(di00222@surrey.ac.uk));
28    %let clientmachine = %nrquote(%nrstr());
29    %let _clientmachine = %nrquote(%nrstr());
30    %let _clientmode = %nrquote(%nrstr(viya));
31    %let sasworklocation="%sysfunc(getoption(work))/";
32    filename _cwd &sasworklocation;
33    data _null_;
34        call symput('_sasworkingdir',pathname('_cwd'));
35    run;
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
      

36    filename _cwd;
NOTE: Fileref _CWD has been deassigned.
37    %let _sasprogramfile = %nrquote(%nrstr());
38    %let _baseurl = %nrquote(%nrstr(https://vfl-040.engage.sas.com/SASStudio/));
39    %let _execenv = %nrquote(%nrstr(SASStudio));
40    %symdel _dataout_mime_type _dataout_name _dataout_url _dataout_table / nowarn;
41    %let _sasws_ = %bquote(%sysfunc(getoption(work)));
42    %let _saswstemp_ = %bquote(%sysfunc(getoption(work)));
43    
44    /* Detect SAS/Graph and setup graph options */
45    data _null_;
46        length rc $255;
47        call symput("graphinit","");
48        call symput("graphterm","");
49        rc=tslvl('sasxgopt','n');
50        _error_=0;
51        if (rc^=' ') then do;
52            call symput("graphinit","goptions reset=all gsfname=_gsfname;");
53            call symput("graphterm","goptions noaccessible;");
54        end;
55    run;
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
      

56    data _null_;
57        length rc 4;
58        rc=sysprod("PRODNUM002");
59        if (rc^=1) then do;
60            call symput("graphinit","");
61            call symput("graphterm","");
62        end;
63    run;
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
      

64    
65    /* Setup ODS destinations */
66    ods _all_ close;
67    %studio_results_directory;
68    filename _htmlout "&_results_prefix_..html";
69    filename _listout "&_results_prefix_..lst";
70    filename _gsfname temp;
71    filename _dataout "&_results_prefix_..dat";
72    ods autonavigate off;
73    ods graphics on;
74    ods html5 (id=web) METATEXT='http-equiv="Content-Security-Policy" content="default-src ''none''; style-src ''unsafe-inline'';
74  ! img-src data: ;"' device=png gpath="&_saswstemp_" path="&_saswstemp_" encoding=utf8 file=_htmlout (title='Results:SAS
74  ! Program.sas') style=Ignite options(bitmap_mode='inline' outline='on' svg_mode='inline' css_prefix=".ods_&SYS_COMPUTE_JOB_ID"
74  ! body_id="div_&SYS_COMPUTE_JOB_ID" );
NOTE: Writing HTML5(WEB) Body file: _HTMLOUT
75    ods listing file=_listout;
76    &graphinit;
77    %studio_initialize_custom_output;
```

```
78   /* endregion */
79   
80   /*****************************************************************
81   * SAS Visual Text Analytics
82   * Sentiment Score Code
83   *
84   * Modify the following macro variables to match your needs.
85   ****************************************************************/
86   
87   /* specifies CAS library information for the CAS table that you would like to score. You must modify the value to provide the
87 ! name of the library that contains the table to be scored. */
88   %let input_caslib_name = "CASUSER";
89   
90   /* specifies the CAS table you would like to score. You must modify the value to provide the name of the input table, such as
90 ! "MyTable". Do not include an extension. */
91   %let input_table_name = "TWEETS";
92   
93   /* specifies the column in the CAS table that contains a unique document identifier. You must modify the value to provide the
93 ! name of the document identifer column in the table. */
94   %let key_column = "ID";
95   
96   /* specifies the column in the CAS table that contains the text data to score. You must modify the value to provide the name of
96 !  the text column in the table. */
97   %let document_column = "Text";
98   
99   /* specifies the CAS library to write the score output tables. You must modify the value to provide the name of the library
99 ! that will contain the output tables that the score code produces. */
100  %let output_caslib_name = "CASUSER";
101  
102  /* specifies the sentiment output CAS table to produce */
103  %let output_sentiment_table_name = "out_sentiment_1";
104  
105  /* specifies the matches output CAS table to produce */
106  %let output_matches_table_name = "out_sent_matches_2";
107  
108  /* specifies the features output CAS table to produce */
109  %let output_features_table_name = "out_sent_features_3";
110  
111  /* specifies the language of the associated SAS Visual Text Analytics project. This should be set automatically to the language
111!  you selected when you created your project. */
112  %let language = "ENGLISH";
113  
114  /* specifies the hostname for the CAS server. This should be set automatically to the host for the associated SAS Visual Text
114! Analytics project. */
115  %let cas_server_hostname = "sas-cas-server-default-client";
116  
117  /* specifies the port for the CAS server. This should be set automatically to the host for the associated SAS Visual Text
117! Analytics project. */
118  %let cas_server_port = 5570;
119  
120  /* creates a session */
121  cas sascas1 host=&cas_server_hostname port=&cas_server_port;
NOTE: The session SASCAS1 connected successfully to Cloud Analytic Services sas-cas-server-default-client using port 5570. The UUID
      is af876e96-a698-f846-8c60-d5d758ccbfb3. The user is di00222@surrey.ac.uk and the active caslib is
      CASUSER(di00222@surrey.ac.uk).
NOTE: The SAS option SESSREF was updated with the value SASCAS1.
NOTE: The SAS macro _SESSREF_ was updated with the value SASCAS1.
NOTE: The session is using 0 workers.
122  libname sascas1 cas sessref=sascas1 datalimit=all;
NOTE: Libref SASCAS1 was successfully assigned as follows:
      Engine:        CAS
      Physical Name: af876e96-a698-f846-8c60-d5d758ccbfb3
123  
124  /* calls the scoring action */
125  proc cas;
126      session sascas1;
127      loadactionset "sentimentAnalysis";
128  
129      action applySent;
130          param
131              table={caslib=&input_caslib_name, name=&input_table_name}
132              docId=&key_column
133              text=&document_column
134              language=&language
135              casOut={caslib=&output_caslib_name, name=&output_sentiment_table_name, replace=TRUE}
136              matchOut={caslib=&output_caslib_name, name=&output_matches_table_name, replace=TRUE}
137              featureOut={caslib=&output_caslib_name, name=&output_features_table_name, replace=TRUE}
138          ;
139      run;
NOTE: Active Session now sascas1.
NOTE: Added action set 'sentimentAnalysis'.
140  quit;
NOTE: The PROCEDURE CAS printed page 1.
NOTE: PROCEDURE CAS used (Total process time):
      real time           0.67 seconds
      cpu time            0.03 seconds
      

141  
142  /* region: Generated postamble */
143  /* Close ODS destinations */
144  &graphterm; ;*';*";*/;run;quit;
145  quit;run;
```

```
146  ods html5 (id=web) close;
147  ods listing close;
148  %if %sysfunc(fileref(_gsfname)) lt 0 %then %do;
149      filename _gsfname clear;
NOTE: Fileref _GSFNAME has been deassigned.
150  %end;
151  %studio_capture_custom_output;
152  /* endregion */
153
```

Results:Sentiment Analysis.sas

Results from sentimentAnalysis.applySent

Output CAS Tables				
CAS Library	Name	Label	Number of Rows	Number of Columns
CASUSER(di00222@surrey.ac.uk)	out_sentiment_1		598	3
CASUSER(di00222@surrey.ac.uk)	out_sent_matches_2		1150	7
CASUSER(di00222@surrey.ac.uk)	out_sent_features_3		0	5

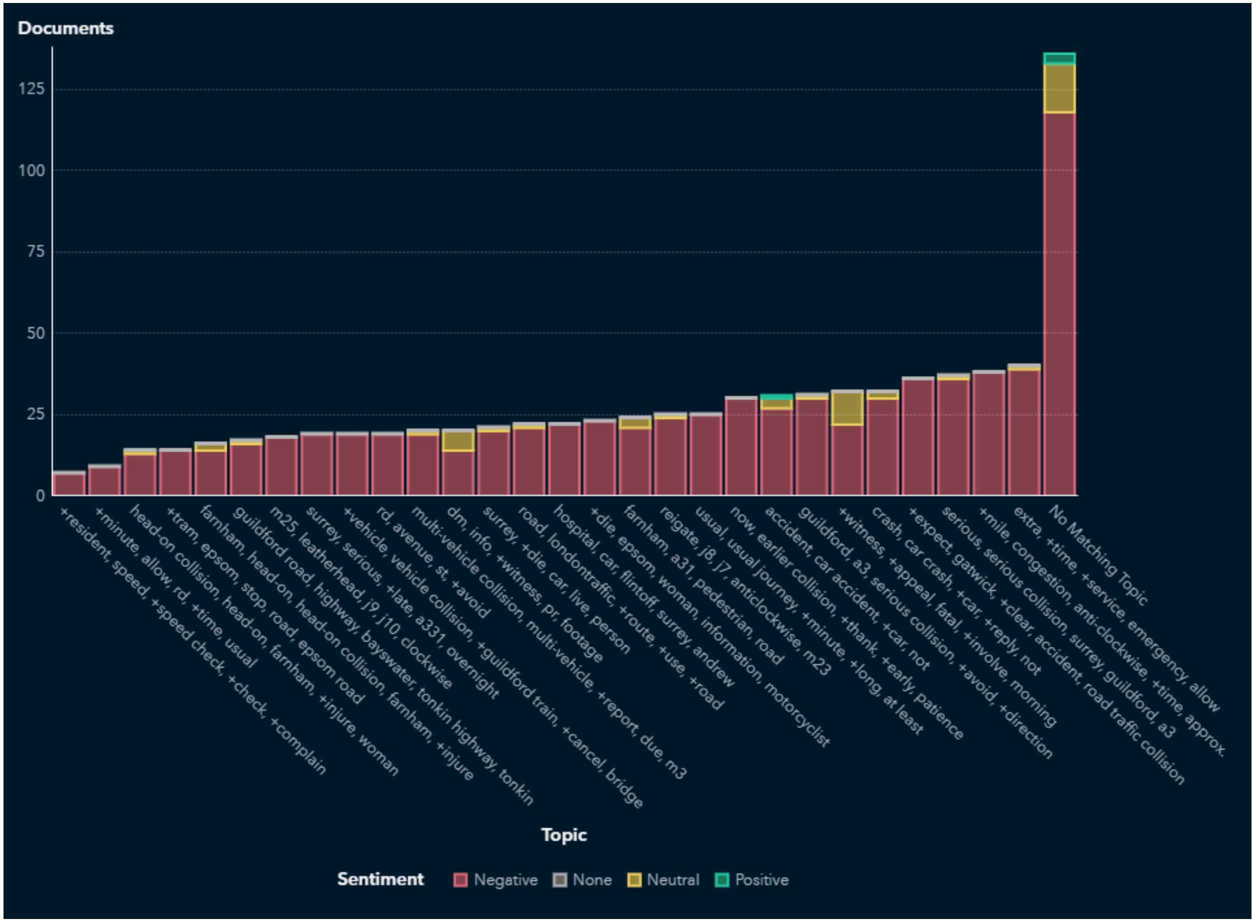
Surrey_Accident_Analysis_Final

"Topics" Results

by: di00222@surrey.ac.uk

Contents

Number of Documents Per Topic



The Number of Documents per Topic report is a visual summary of the topics in the data. 28 of the bars represent the number of topics identified. Bar 29 in the chart represents the documents that have not been assigned to any topic. The height of each bar represents the number of documents placed in each topic. Documents may be placed into more than a single topic.

Applying this set of topics to this data set results in 136 (22.74%) of the documents not being placed in any topic. If more coverage of the data is desired, the cutoff for documents and/or terms can be lowered, or more topics can be added, or existing topics expanded in scope.

If a Sentiment node is placed before the Topics node in the pipeline, each bar representing a topic is split between positive, neutral, negative, and no sentiment by color.

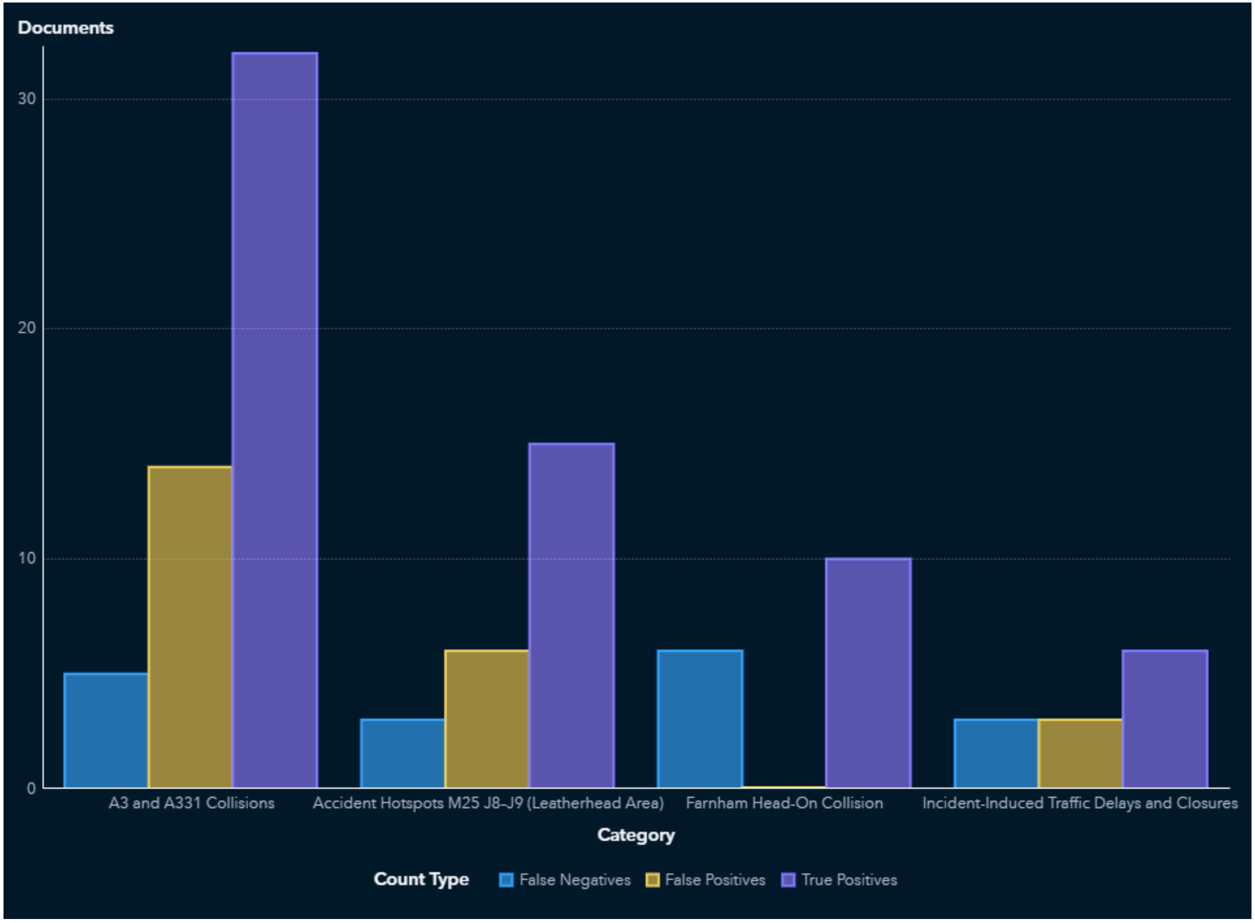
Surrey_Accident_Analysis_Final

"Categories" Results

by: di00222@surrey.ac.uk

Contents

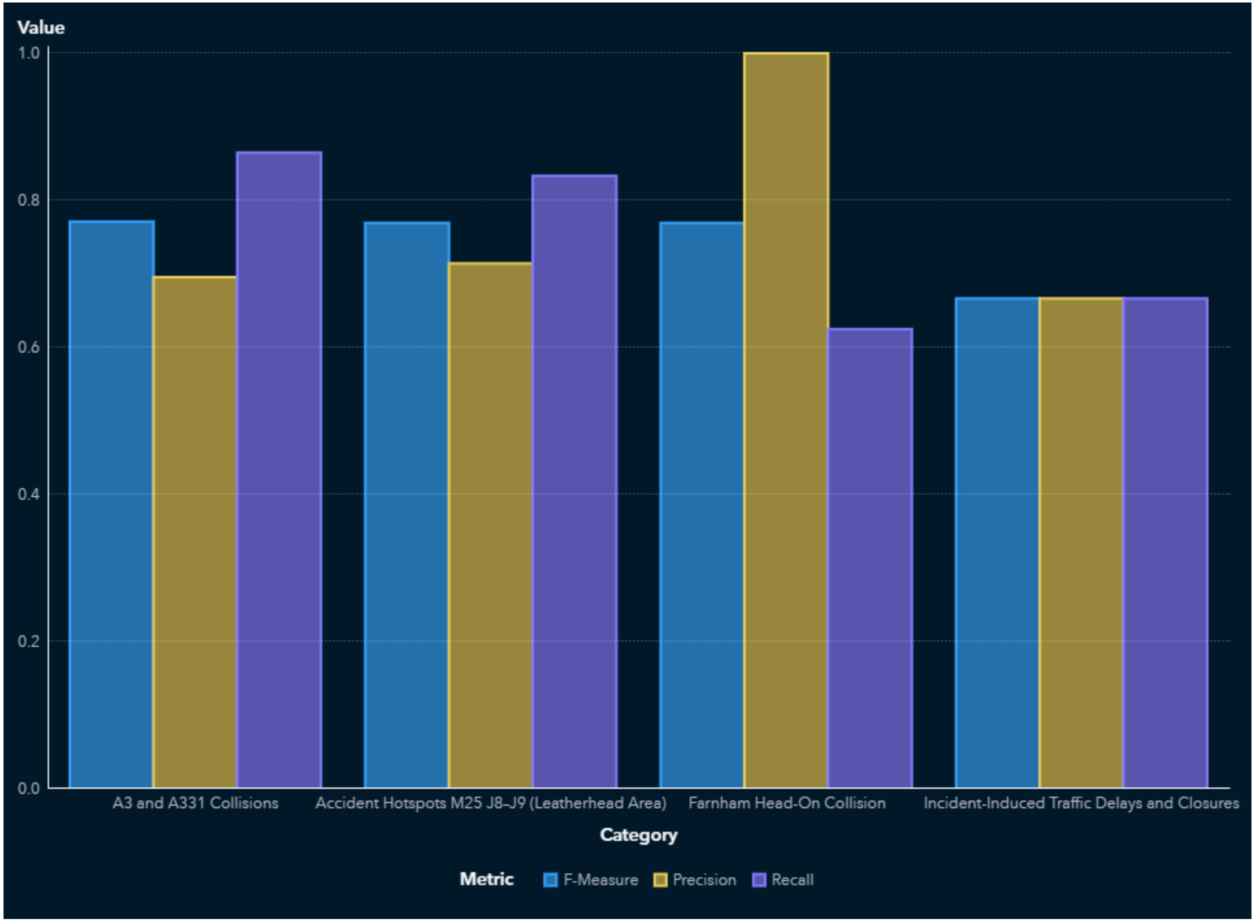
Diagnostic Counts for Automatically Generated Categories



The Diagnostic Counts chart depicts the number of documents matched by a category that are true positives, false positives, and false negatives.   The chart enables the model developer to target weaknesses in a particular category. For example, the category with the largest count of false positives is "A3 and A331 Collisions". To improve this category, narrow the scope of the rule definition to reduce the number of matches that are incorrect. Narrowing can be achieved by adding more specific operators, such as SENT, DIST_n, NOTIN, or MINOC_n, or by evaluating keywords in the category definition to determine their reliability.

The category with the largest count of false negatives is "Farnham Head-On Collision". To improve this category, broaden or deepen the scope of the category definition. To broaden the scope, use broader operators, such as AND, OR, NOT, or MAXSENT_n. To deepen the scope of the category, use additional operators and keywords.

Diagnostic Metrics for Automatically Generated Categories



The Diagnostic Metrics chart depicts calculations of recall, precision, and f-score for each category. These calculations are based upon the counts in the Diagnostic Counts chart. In this data set the best performing category, based upon f-score, is "A3 and A331 Collisions". An f-score of 0.77 indicates the model is performing at a reasonably good level of accuracy.

F-score depicts a balance between recall and precision. Recall is a score that focuses on the optimization of true positives vs. false negatives and gives insight into the question of how many documents were missed by the category definition, while precision is a score that focuses on the optimization of true positives vs. false positives and gives insight into the question of how many extra documents were matched beyond the targeted documents.

This model needs more work on precision. To improve precision, try examining the matched documents to determine how the rule definition can be modified to exclude the ones that should not be matched. Start with use of narrower operators such as SENT, DIST_n, NOTIN, or MINOC_n instead of AND, OR, and NOT.