
NLP

Introduction to NLP

Evaluation of IR

Evaluation

- Size of index
- Speed of indexing
- Speed of retrieval
- Accuracy
- Timeliness
- Ease of use
- Expressiveness of search language

Contingency Table

	retrieved	not retrieved	
relevant	$w=tp$	$x=fn$	$n_1 = w + x$
not relevant	$y=fp$	$z=tn$	
	$n_2 = w + y$		N

Precision and Recall

Recall: $\frac{w}{w+x}$

Precision: $\frac{w}{w+y}$

Issues

- Why not use accuracy A=(w+z)/N?
- Average precision
- Report when P=R
- F measure:
  - $F=(\beta^2+1)PR/(\beta^2P+R)$
- F1 measure:
  - F1 = 2/(1/R+1/P) : harmonic mean of P and R

Sample TREC query

<top>
<num> Number: 305
<title> Most Dangerous Vehicles

<desc> Description:
Which are the most crashworthy, and least crashworthy, passenger vehicles?

<narr> Narrative:
A relevant document will contain information on the crashworthiness of a given vehicle or vehicles that can be used to draw a comparison with other vehicles. The document will have to describe/compare vehicles, not drivers. For instance, it should be expected that vehicles preferred by 16-25 year-olds would be involved in more crashes, because that age group is involved in more crashes. I would view number of fatalities per 100 crashes to be more revealing of a vehicle's crashworthiness than the number of crashes per 100,000 miles, for example.
</top>

LA031689-0177
FT922-1008
LA090190-0126
LA101190-0218
LA082690-0158
LA112590-0109
FT944-136
LA020590-0119
FT944-5300
LA052190-0048
LA051689-0139
FT944-9371
LA032390-0172
LA042790-0172
LA021790-0136
LA092289-0167
LA111189-0013
LA120189-0179
LA020490-0021
LA122989-0063
LA091389-0119
LA072189-0048
FT944-15615
LA091589-0101
LA021289-0208

<DOCNO> LA031689-0177 </DOCNO>
<DOCID> 31701 </DOCID>
<DATE><P>March 16, 1989, Thursday, Home Edition </P></DATE>
<SECTION><P>Business; Part 4; Page 1; Column 5; Financial Desk </P></SECTION>
<LENGTH><P>586 words </P></LENGTH>
<HEADLINE><P>AGENCY TO LAUNCH STUDY OF FORD BRONCO II AFTER HIGH RATE OF ROLL-OVER ACCIDENTS </P></HEADLINE>
<BYLINE><P>By LINDA WILLIAMS, Times Staff Writer </P></BYLINE>
<TEXT>
<P>The federal government's highway safety watchdog said Wednesday that the Ford Bronco II appears to be involved in more fatal roll-over accidents than other vehicles in its class and that it will seek to determine if the vehicle itself contributes to the accidents. </P>
<P>The decision to do an engineering analysis of the Ford Motor Co. utility-sport vehicle grew out of a federal accident study of the Suzuki Samurai, said Tim Hurd, a spokesman for the National Highway Traffic Safety Administration. NHTSA looked at Samurai accidents after Consumer Reports magazine charged that the vehicle had basic design flaws. </P>
<P>Several Fatalities </P>
<P>However, the accident study showed that the "Ford Bronco II appears to have a higher number of single-vehicle, first event roll-overs, particularly those involving fatalities," Hurd said. The engineering analysis of the Bronco, the second of three levels of investigation conducted by NHTSA, will cover the 1984-1989 Bronco II models, the agency said. </P>
<P>According to a Fatal Accident Reporting System study included in the September report on the Samurai, 43 Bronco II single-vehicle roll-overs caused fatalities, or 19 of every 100,000 vehicles. There were eight Samurai fatal roll-overs, or 6 per 100,000; 13 involving the Chevrolet S10 Blazers or GMC Jimmy, or 6 per 100,000, and six fatal Jeep Cherokee roll-overs, for 2.5 per 100,000. After the accident report, NHTSA declined to investigate the Samurai. </P>
...
</TEXT>
<GRAPHIC><P> Photo, The Ford Bronco II "appears to have a higher number of single-vehicle, first event roll-overs," a federal official said. </P></GRAPHIC>
<SUBJECT>
<P>TRAFFIC ACCIDENTS; FORD MOTOR CORP; NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION; VEHICLE INSPECTIONS; RECREATIONAL VEHICLES; SUZUKI MOTOR CO; AUTOMOBILE SAFETY </P>
</SUBJECT>
</DOC>

TREC (cont'd)

- http://trec.nist.gov/tracks.html
- http://trec.nist.gov/presentations/presentations.html

Most Used Reference Collections

- Generic retrieval
  - OHSUMED, CRANFIELD, CACM
- Text classification
  - Reuters, 20newsgroups
- Question answering
  - TREC-QA
- Web
  - DOTGOV, wt100g
- Blogs
  - Buzzmetrics datasets
- TREC ad hoc collections, 2-6 GB
- TREC Web collections, 2-100GB

Comparing Two Systems

- Comparing A and B
- One query?
- Average performance?
- Need: A to consistently outperform B

[Example from James Allan]

The Sign Test

- Example 1:
  - A > B (12 times)
  - A = B (25 times)
  - A < B (3 times)
  - $p < 0.035$ (significant at the 5% level)
- Example 2:
  - A > B (18 times)
  - A < B (9 times)
  - $p < 0.122$ (not significant at the 5% level)
- External link:
  - http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html

Other Tests

- Student t-test: takes into account the actual performances, not just which system is better
  - http://www.fon.hum.uva.nl/Service/Statistics/Student_t_Test.html
  - http://www.socialresearchmethods.net/kb/stat_t.php
- Wilcoxon Matched-Pairs Signed-Ranks Test
  - http://www.fon.hum.uva.nl/Service/Statistics/Signed_Rank_Test.html


NLP