GTE @ NTCIR GeoTime

"Experiments with Geo-Temporal Expressions filtering and query expansion at document and phrase context"

Jorge Machado

jmachado@estgp.pt

Outline

- Challenges
- Experiment Overview
- Collection Processing and Statistics
- Indexes
- Topics Processing
- Experimented RUNS
- Results
- Future Work

LGTE – Geo-Temporal Retrieval

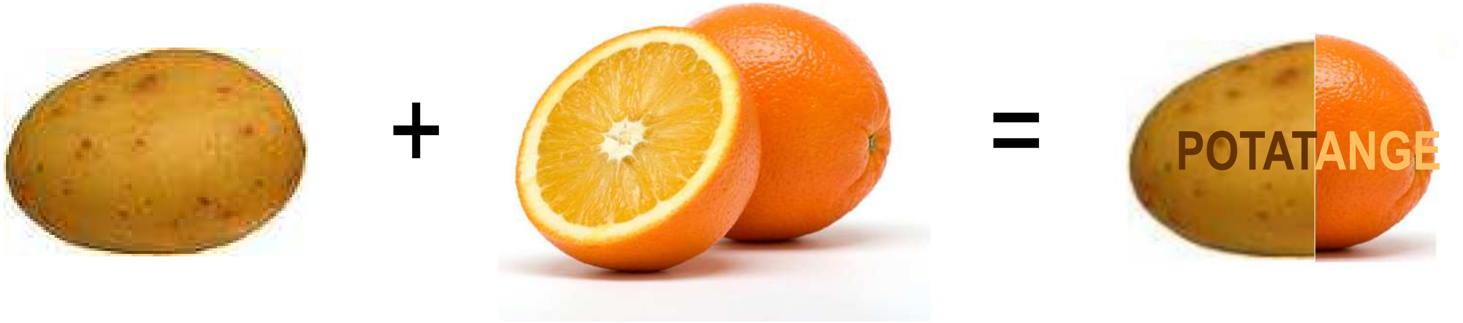
Geo-Temporal semantic layer Challenge:

Combination of standard Information Retrieval (IR) mechanisms with new techniques for addressing the geographic and temporal dimensions of relevance.

Score Functions Challenge

- How to score documents using Geographic and Temporal dimensions in one unique Scale?

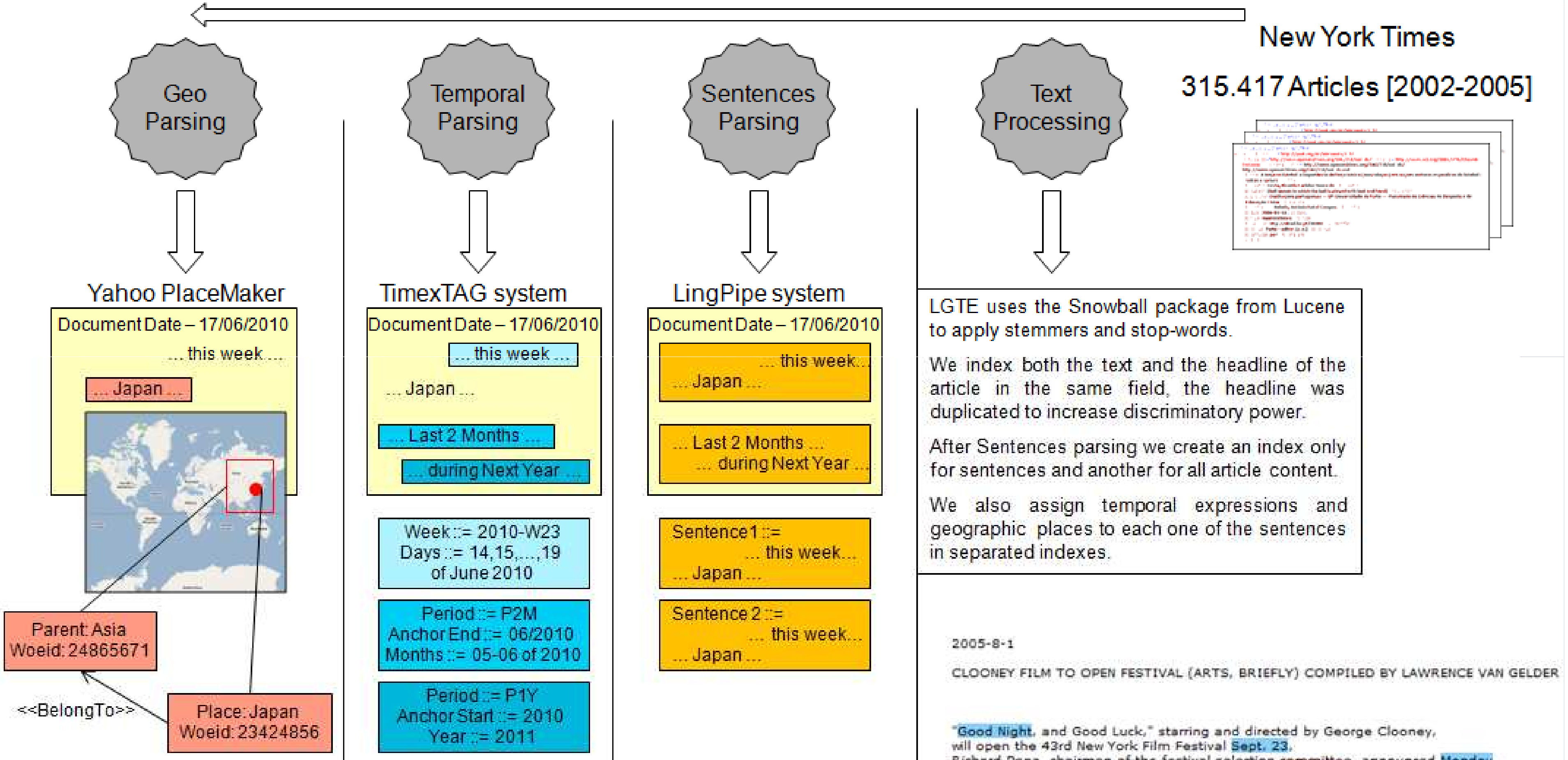
… How can we sum this?



Experiment Setup

- We combined scores … if there was nothing more that we can do.
- We created annotations for the documents and the topics.
- We used the annotations in the topics to filter the documents from the result set.
- We experiment several context levels:
  - Text
  - Sentence
- We used a mix approach combine scores from sentences and text (usual in book search).

Document processing



LGTE uses the Snowball package from Lucene to apply stemmers and stop-words.

We index both the text and the headline of the article in the same field, the headline was duplicated to increase discriminatory power.

After Sentences parsing we create an index only for sentences and another for all article content.

We also assign temporal expressions and geographic places to each one of the sentences in separated indexes.

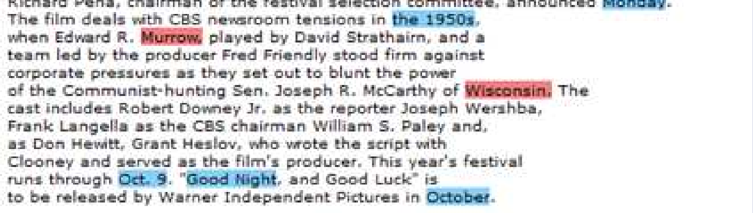
Yahoo PlaceMaker recognize and disambiguate the names of places and temporal expressions given in the text, also assigning documents to the encompassing geo-temporal scopes that they discuss as a whole.

TimexTag recognize temporal expressions and mark them with Timex2 tags. We normalize some of those tags to Dates of the format: YYYY[MM[DD]]

LingPipe provide a tool to split the text into sentences. We obtained more than 11 million sentences. Our purpose was to discriminate fine grained contextualized information.

2005-8-1

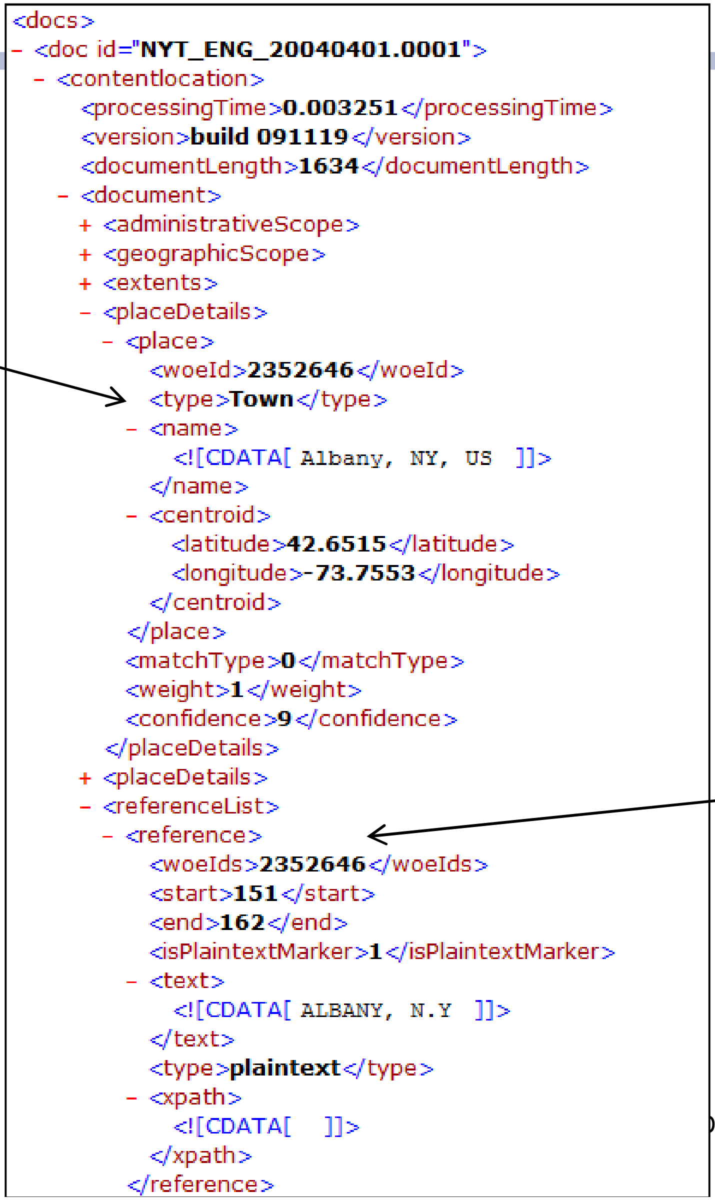
CLOONEY FILM TO OPEN FESTIVAL (ARTS, BRIEFLY) COMPILED BY LAWRENCE VAN GELDER



"Good Night, and Good Luck," starring and directed by George Clooney,
will open the 43rd New York Film Festival Sept. 23,
Richard Pena, chairman of the festival selection committee, announced Monday.
The film deals with CBS newsroom tensions in the 1950s,
when Edward R. Murrow, played by David Strathairn, and a
team led by the producer Fred Friendly stood firm against
corporate pressures as they set out to blunt the power
of the Communist-hunting Sen. Joseph R. McCarthy of Wisconsin. The
cast includes Robert Downey Jr. as the reporter Joseph Wershba,
Frank Langella as the CBS chairman William S. Paley and,
as Don Hewitt, Grant Heslov, who wrote the script with
Clooney and served as the film's producer. This year's festival
runs through Oct. 9. "Good Night, and Good Luck" is
to be released by Warner Independent Pictures in October.

Geo-Parsed Document Example using Yahoo PlaceMaker

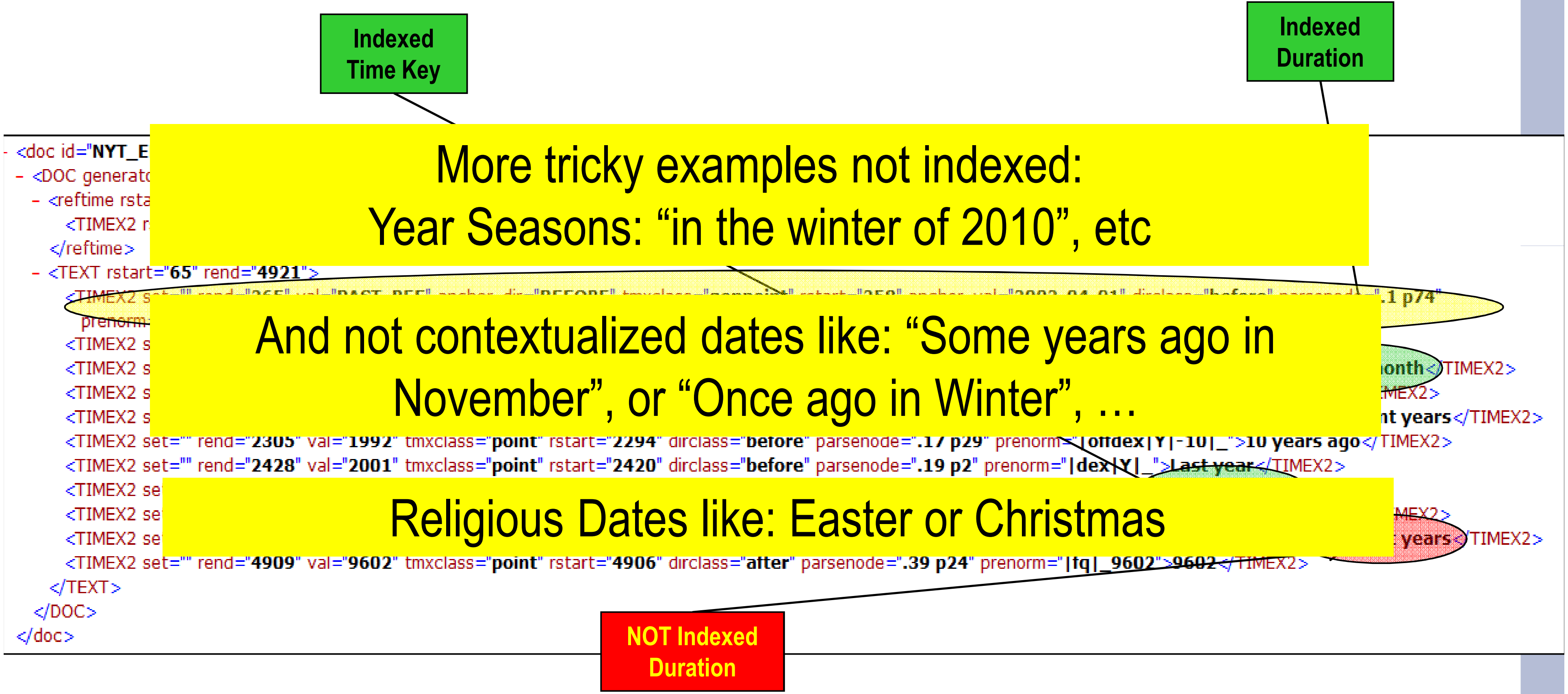
Indexed PlaceType



```
<docs>
- <doc id="NYT_ENG_20040401.0001">
  - <contentlocation>
      <processingTime>0.003251</processingTime>
      <version>build 091119</version>
      <documentLength>1634</documentLength>
    - <document>
      + <administrativeScope>
      + <geographicScope>
      + <extents>
      - <placeDetails>
        - <place>
            <woeId>2352646</woeId>
            <type>Town</type>
          - <name>
              <![CDATA[ Albany, NY, US  ]]>
            </name>
          - <centroid>
              <latitude>42.6515</latitude>
              <longitude>-73.7553</longitude>
            </centroid>
          </place>
          <matchType>0</matchType>
          <weight>1</weight>
          <confidence>9</confidence>
        </placeDetails>
      + <placeDetails>
      - <referenceList>
        - <reference>
            <woeIds>2352646</woeIds>
            <start>151</start>
            <end>162</end>
            <isPlaintextMarker>1</isPlaintextMarker>
          - <text>
              <![CDATA[ ALBANY, N.Y  ]]>
            </text>
            <type>plaintext</type>
          - <xpath>
              <![CDATA[   ]]>
            </xpath>
          </reference>
```

Indexed Place

Temporal Parsed Document using TimexTAG



Indexed Time Key

Indexed Duration

More tricky examples not indexed:
Year Seasons: “in the winter of 2010”, etc

And not contextualized dates like: “Some years ago in November”, or “Once ago in Winter”, …

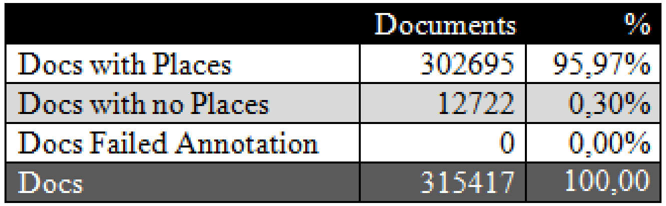
Religious Dates like: Easter or Christmas

NOT Indexed Duration

```
- <doc id="NYT_E
  - <DOC generato
    - <reftime rsta
        <TIMEX2 r
      </reftime>
    - <TEXT rstart="65" rend="4921">
        <TIMEX2 set="" rend="265" val="PAST_REF" anchor_dir="BEFORE" tmxclass="genpoint" rstart="258" anchor_val="2002-04-01" dirclass="before" parsenode=".1 p74"
          prenorm
        <TIMEX2 s
        <TIMEX2 s                                          onth</TIMEX2>
        <TIMEX2 s                                          MEX2>
        <TIMEX2 s                                          nt years</TIMEX2>
        <TIMEX2 set="" rend="2305" val="1992" tmxclass="point" rstart="2294" dirclass="before" parsenode=".17 p29" prenorm="|offdex|Y|-10|_">10 years ago</TIMEX2>
        <TIMEX2 set="" rend="2428" val="2001" tmxclass="point" rstart="2420" dirclass="before" parsenode=".19 p2" prenorm="|dex|Y|_">Last year</TIMEX2>
        <TIMEX2 se
        <TIMEX2 se                                         MEX2>
        <TIMEX2 se                                         years</TIMEX2>
        <TIMEX2 set="" rend="4909" val="9602" tmxclass="point" rstart="4906" dirclass="after" parsenode=".39 p24" prenorm="|fq|_9602">9602</TIMEX2>
      </TEXT>
    </DOC>
</doc>
```

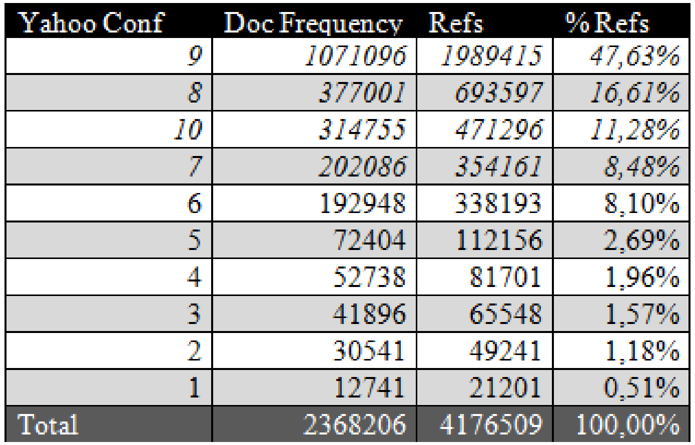
Geo-Parsing Statistics

Table 1 – Geo-Parsing General Statistics.



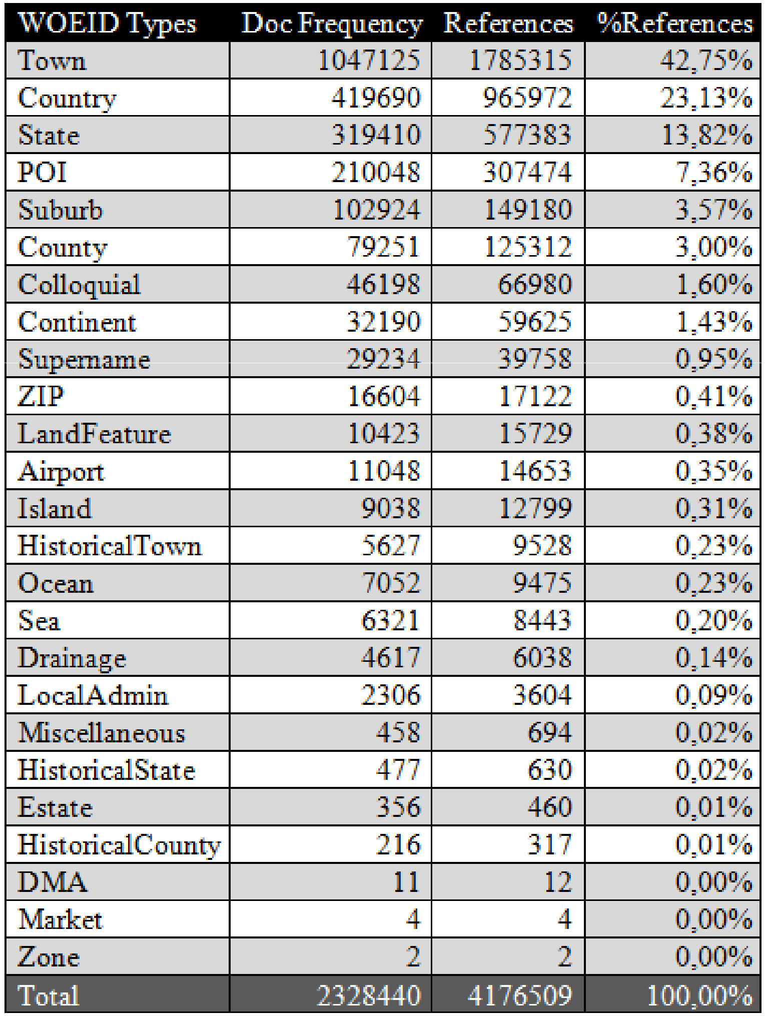
	Documents	%
Docs with Places	302695	95,97%
Docs with no Places	12722	0,30%
Docs Failed Annotation	0	0,00%
Docs	315417	100,00

Table 3 – Yahoo Place Maker confidence degree.



Yahoo Conf	Doc Frequency	Refs	% Refs
9	1071096	1989415	47,63%
8	377001	693597	16,61%
10	314755	471296	11,28%
7	202086	354161	8,48%
6	192948	338193	8,10%
5	72404	112156	2,69%
4	52738	81701	1,96%
3	41896	65548	1,57%
2	30541	49241	1,18%
1	12741	21201	0,51%
Total	2368206	4176509	100,00%

Table 2 – Place types distribution over documents.



WOEID Types	Doc Frequency	References	%References
Town	1047125	1785315	42,75%
Country	419690	965972	23,13%
State	319410	577383	13,82%
POI	210048	307474	7,36%
Suburb	102924	149180	3,57%
County	79251	125312	3,00%
Colloquial	46198	66980	1,60%
Continent	32190	59625	1,43%
Supername	29234	39758	0,95%
ZIP	16604	17122	0,41%
LandFeature	10423	15729	0,38%
Airport	11048	14653	0,35%
Island	9038	12799	0,31%
HistoricalTown	5627	9528	0,23%
Ocean	7052	9475	0,23%
Sea	6321	8443	0,20%
Drainage	4617	6038	0,14%
LocalAdmin	2306	3604	0,09%
Miscellaneous	458	694	0,02%
HistoricalState	477	630	0,02%
Estate	356	460	0,01%
HistoricalCounty	216	317	0,01%
DMA	11	12	0,00%
Market	4	4	0,00%
Zone	2	2	0,00%
Total	2328440	4176509	100,00%

Temporal Parsing Statistics

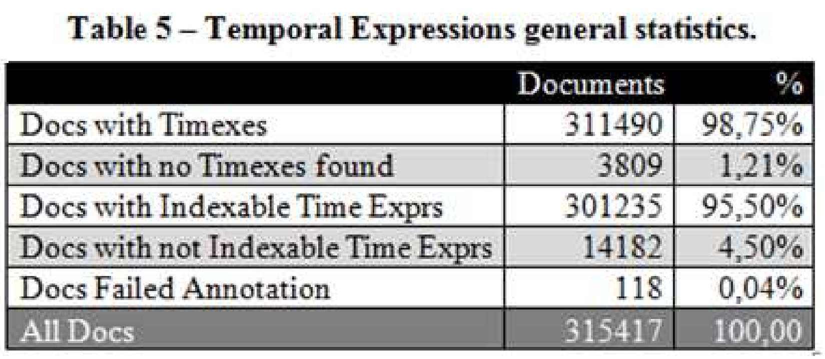


Table 5 – Temporal Expressions general statistics.

	Documents	%
Docs with Timexes	311490	98,75%
Docs with no Timexes found	3809	1,21%
Docs with Indexable Time Exprs	301235	95,50%
Docs with not Indexable Time Exprs	14182	4,50%
Docs Failed Annotation	118	0,04%
All Docs	315417	100,00

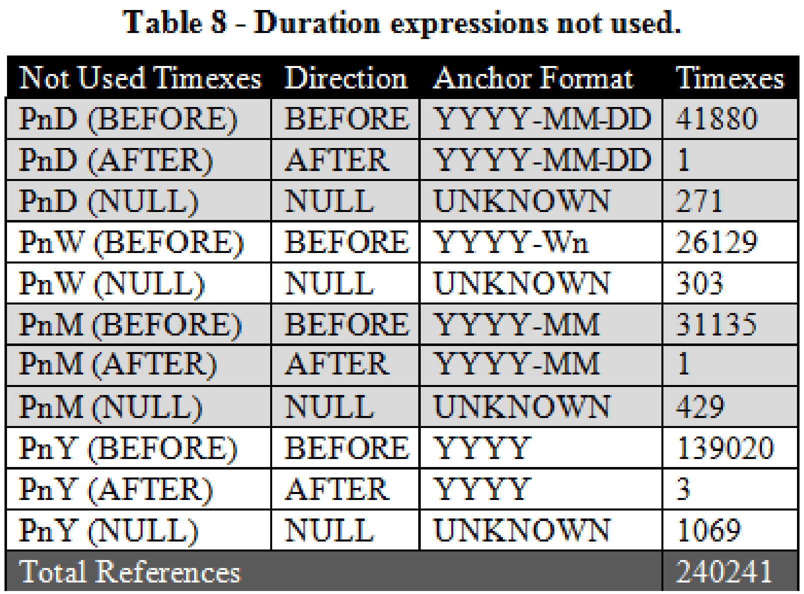


Table 8 - Duration expressions not used.

Not Used Timexes	Direction	Anchor Format	Timexes
PnD (BEFORE)	BEFORE	YYYY-MM-DD	41880
PnD (AFTER)	AFTER	YYYY-MM-DD	1
PnD (NULL)	NULL	UNKNOWN	271
PnW (BEFORE)	BEFORE	YYYY-Wn	26129
PnW (NULL)	NULL	UNKNOWN	303
PnM (BEFORE)	BEFORE	YYYY-MM	31135
PnM (AFTER)	AFTER	YYYY-MM	1
PnM (NULL)	NULL	UNKNOWN	429
PnY (BEFORE)	BEFORE	YYYY	139020
PnY (AFTER)	AFTER	YYYY	3
PnY (NULL)	NULL	UNKNOWN	1069
Total References			240241

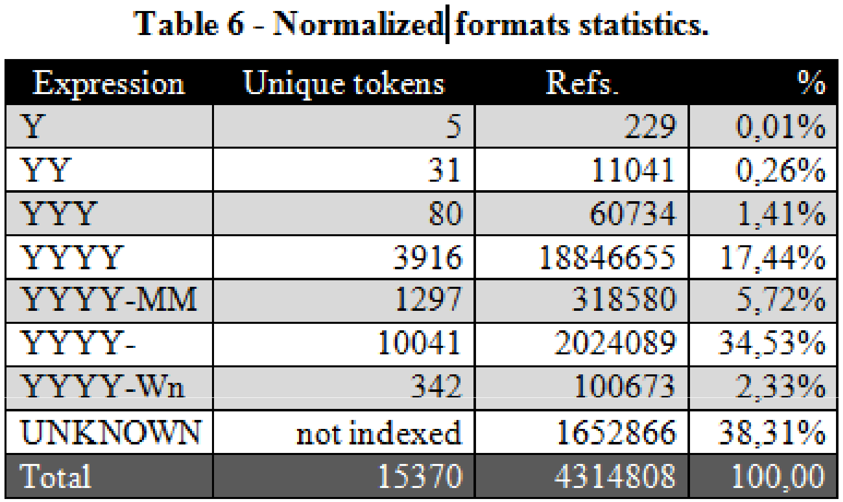


Table 6 - Normalized formats statistics.

Expression	Unique tokens	Refs.	%
Y	5	229	0,01%
YY	31	11041	0,26%
YYY	80	60734	1,41%
YYYY	3916	18846655	17,44%
YYYY-MM	1297	318580	5,72%
YYYY-	10041	2024089	34,53%
YYYY-Wn	342	100673	2,33%
UNKNOWN	not indexed	1652866	38,31%
Total	15370	4314808	100,00

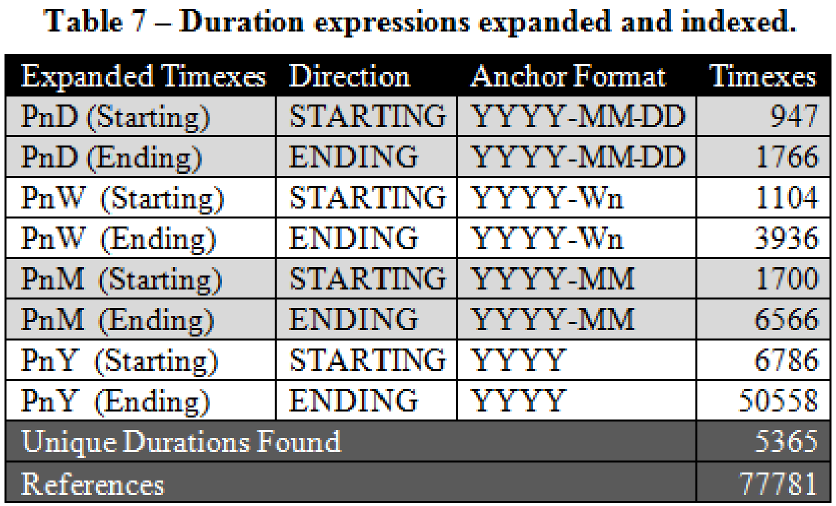


Table 7 – Duration expressions expanded and indexed.

Expanded Timexes	Direction	Anchor Format	Timexes
PnD (Starting)	STARTING	YYYY-MM-DD	947
PnD (Ending)	ENDING	YYYY-MM-DD	1766
PnW (Starting)	STARTING	YYYY-Wn	1104
PnW (Ending)	ENDING	YYYY-Wn	3936
PnM (Starting)	STARTING	YYYY-MM	1700
PnM (Ending)	ENDING	YYYY-MM	6566
PnY (Starting)	STARTING	YYYY	6786
PnY (Ending)	ENDING	YYYY	50558
Unique Durations Found			5365
References			77781

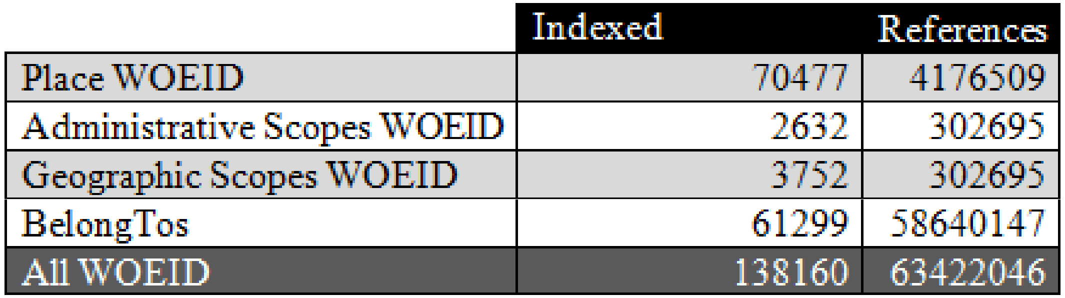
Indexes

Index Name	TEXT	SENTENCES
Terms	2*HEADLINE+ 1*TEXT	Only the TEXT of sentence
Places	WOEID’s	… in sentence
BelongTos	Places Ancestors	“
PlaceType	Types of Places	“
Dates	Normalized Timexes	“
Durations	Normalized Duration Timexes	“
DatesAndDurations	All Normalized Timexes	“
DateType	Types of Dates (exact, month, year)	“

Indexed Entities

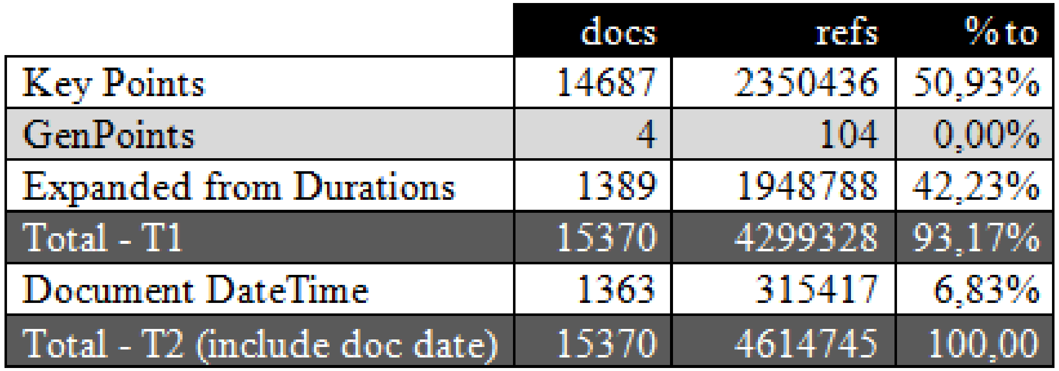
315.417 documents containing 11.702.480 Sentences

Table 4 - Normalized WOEID's.



	Indexed	References
Place WOEID	70477	4176509
Administrative Scopes WOEID	2632	302695
Geographic Scopes WOEID	3752	302695
BelongTos	61299	58640147
All WOEID	138160	63422046

Table 9 – Indexed Temporal Expressions



	docs	refs	% to
Key Points	14687	2350436	50,93%
GenPoints	4	104	0,00%
Expanded from Durations	1389	1948788	42,23%
Total - T1	15370	4299328	93,17%
Document DateTime	1363	315417	6,83%
Total - T2 (include doc date)	15370	4614745	100,00

Topics Processing

GeoTemporal Expressions … Filter with or Search For ??

→ We choose to Filter to in order to minimize dimension scores combination

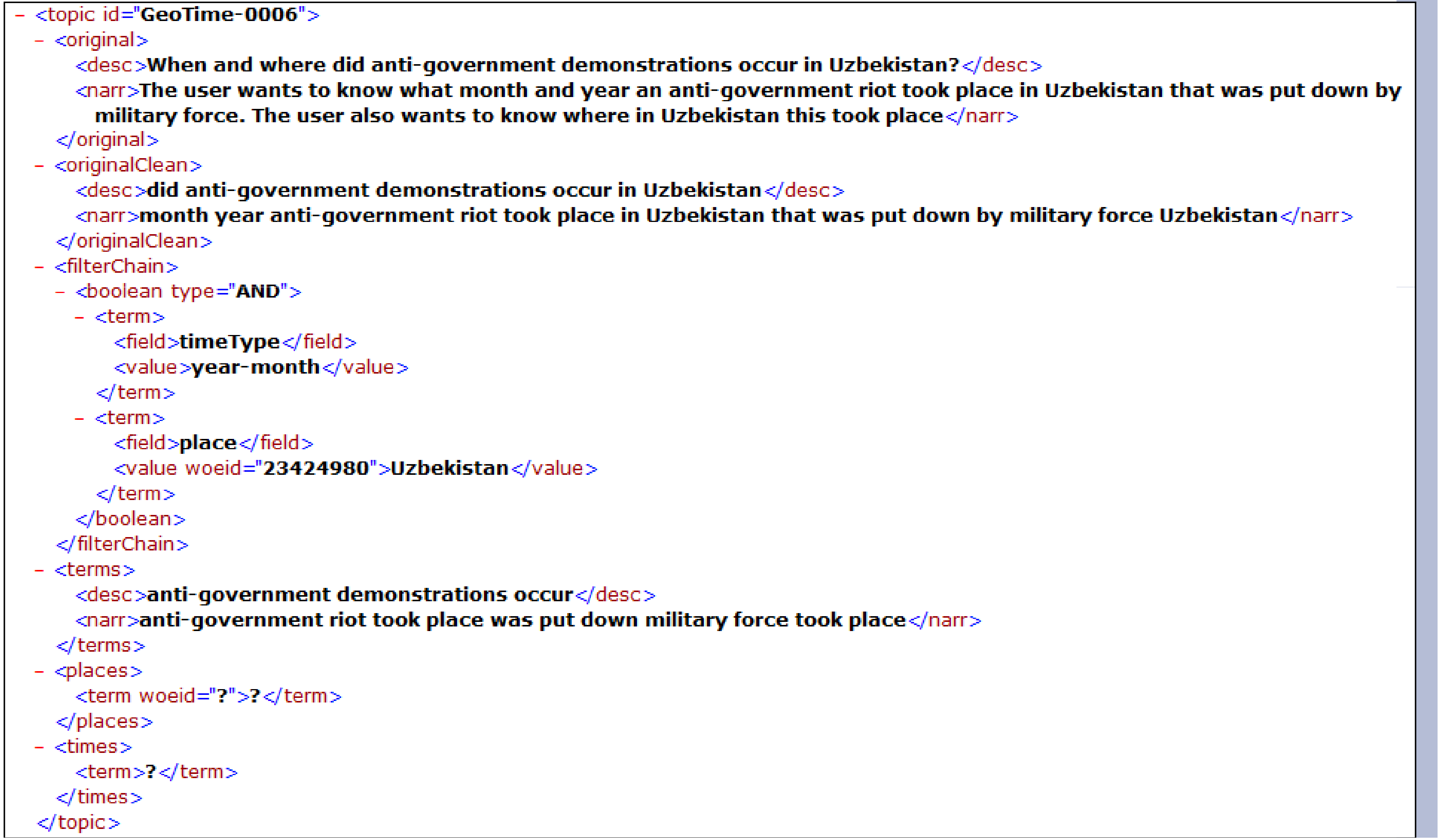
Filter Selection Rules:

- the set of all of the geographic and temporal expressions which occur near an adverb like “what”, “where”, “when”, or the compositions “How long after/before”.
  e.g. “In **what** *city*", "In **what province** of *China*", “**How long after** …”
- “Users Needs” expressed with: “wants to find”, “would like to know", "which one“.
  e.g. “…**wants to know** what *month and year*…”, “**want to know** the *country*”, “**want to know** the *exact date*”

Filters:

- **Places**
  - WOEID’s mostly using ***belongTos*** **index** witch includes the WOEID’s and the ancestors for the found WOEID’s.
- **Types of places**
  - Found in queries: **city, province, country**
- **Temporal expressions**
  - All normalized temporal annotations of the type format **YYYY[MM[DD]]** - mapped dates obtained from time periods (e.g. the last two months), and dates obtained from time keys (e.g. Yesterday(Anchor: 12-05-2005), 13 January(Anchor:15-01-2002) or 15-04-2010)
- **Types of temporal expressions**
  - Found in queries: **exact-date, year, year-month any**

Processed Topic Example



```
- <topic id="GeoTime-0006">
  - <original>
      <desc>When and where did anti-government demonstrations occur in Uzbekistan?</desc>
      <narr>The user wants to know what month and year an anti-government riot took place in Uzbekistan that was put down by
        military force. The user also wants to know where in Uzbekistan this took place</narr>
    </original>
  - <originalClean>
      <desc>did anti-government demonstrations occur in Uzbekistan</desc>
      <narr>month year anti-government riot took place in Uzbekistan that was put down by military force Uzbekistan</narr>
    </originalClean>
  - <filterChain>
    - <boolean type="AND">
      - <term>
          <field>timeType</field>
          <value>year-month</value>
        </term>
      - <term>
          <field>place</field>
          <value woeid="23424980">Uzbekistan</value>
        </term>
      </boolean>
    </filterChain>
  - <terms>
      <desc>anti-government demonstrations occur</desc>
      <narr>anti-government riot took place was put down military force took place</narr>
    </terms>
  - <places>
      <term woeid="?">?</term>
    </places>
  - <times>
      <term>?</term>
    </times>
  </topic>
```

Experimental RUNS

RUN	Filter Granularity	Filter Documents without Temporal OR Geo Expressions	Filter or Pseudo Relevance Feedback Query Expansion	Score Function Granularity
05	Text	Yes	-	Text
04	Sentence	Yes	Yes	Sentence
03	Text	Yes	-	Text
02	Sentence	Yes	Yes	Sentence
01	Text	Yes	Yes	Combination

Score Functions (Constant in all runs)

Simple Linear Combination of Index

$\text{Score}(q,d) = \alpha * \text{bm25Text}(q,d) + \beta * \text{GeoScore}(q,d) + \gamma * \text{TimeScore}(q,d)$

For document d and query q, with $\alpha, \beta, \gamma = 1$

(Scores used for Places or Dates that were not considered Filters)

- GeoScore

$0.7*\text{bm25}_{\text{places}}(d_{\text{places}},q_{\text{places}}) + 0.3*\text{bm25}_{\text{belongTos}}(d_{\text{belongTos}},q_{\text{belongTos}})$

- TimeScore

$0.7*\text{bm25dates}(d_{\text{dates}},q_{\text{dates}}) + 0.3*\text{bm25}_{\text{durations}}(d_{\text{durations}},q_{\text{durations}})$

Pseudo Relevance Feedback (PRF) Query Expansion (QE)

- Rocchio Algorithm adapted for multiple fields scoring.
  - Base Formula

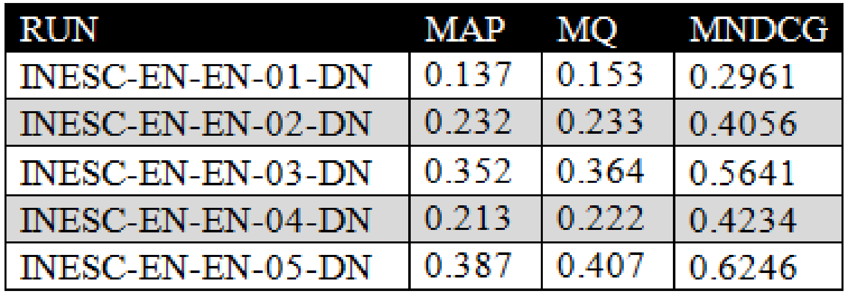
$$q_{i+1} = \alpha \cdot q_i + \frac{\beta}{|D|} \cdot \sum_{d_r \in D} termWeight(d_r)$$

  - Fields (The same set of fields used in Score function)
  - Details of in:
    - [Jorge Machado, Bruno Martins and José Borbinha. Experiments with N-Gram Prefixes on a Multinomial Language Model versus Lucene's off-the-shelf ranking scheme and Rocchio Query Expansion (TEL@CLEF Monolingual Task). ECDL/CLEF, Corfu, Greece, 2009.]

Results

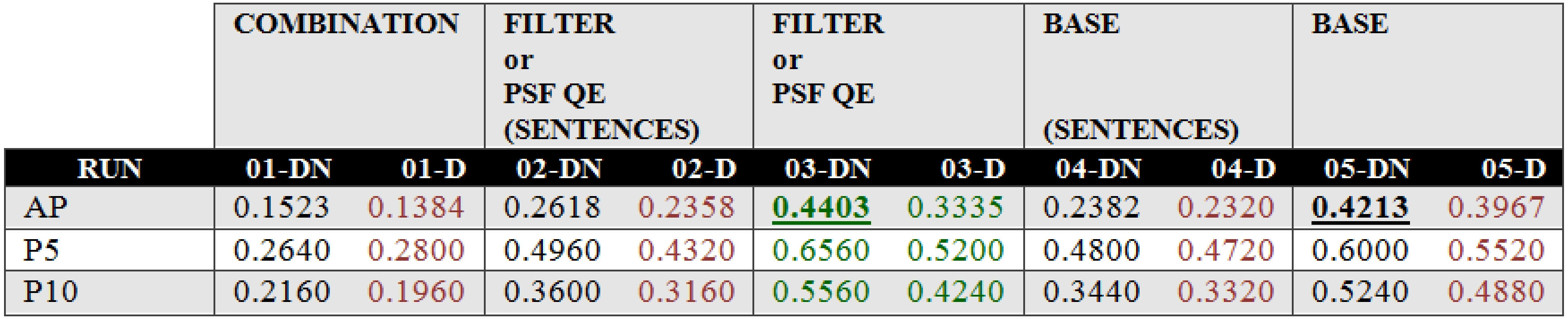
Official Results

Wrong Results
For 01-DN
And
03-DN

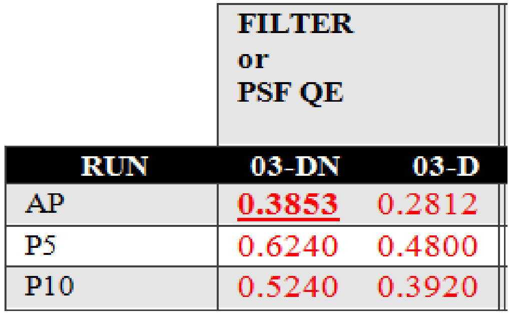


RUN	MAP	MQ	MNDCG
INESC-EN-EN-01-DN	0.137	0.153	0.2961
INESC-EN-EN-02-DN	0.232	0.233	0.4056
INESC-EN-EN-03-DN	0.352	0.364	0.5641
INESC-EN-EN-04-DN	0.213	0.222	0.4234
INESC-EN-EN-05-DN	0.387	0.407	0.6246

Treceval results using binary relevance (significance test from 05-DN to 03-DN = 0.0539)



	COMBINATION		FILTER or PSF QE (SENTENCES)		FILTER or PSF QE		BASE (SENTENCES)		BASE	
RUN	01-DN	01-D	02-DN	02-D	03-DN	03-D	04-DN	04-D	05-DN	05-D
AP	0.1523	0.1384	0.2618	0.2358	0.4403	0.3335	0.2382	0.2320	0.4213	0.3967
P5	0.2640	0.2800	0.4960	0.4320	0.6560	0.5200	0.4800	0.4720	0.6000	0.5520
P10	0.2160	0.1960	0.3600	0.3160	0.5560	0.4240	0.3440	0.3320	0.5240	0.4880



	FILTER or PSF QE	
RUN	03-DN	03-D
AP	0.3853	0.2812
P5	0.6240	0.4800
P10	0.5240	0.3920

* Wrong results reported
in the poster for run
03-DN

Detected Problems

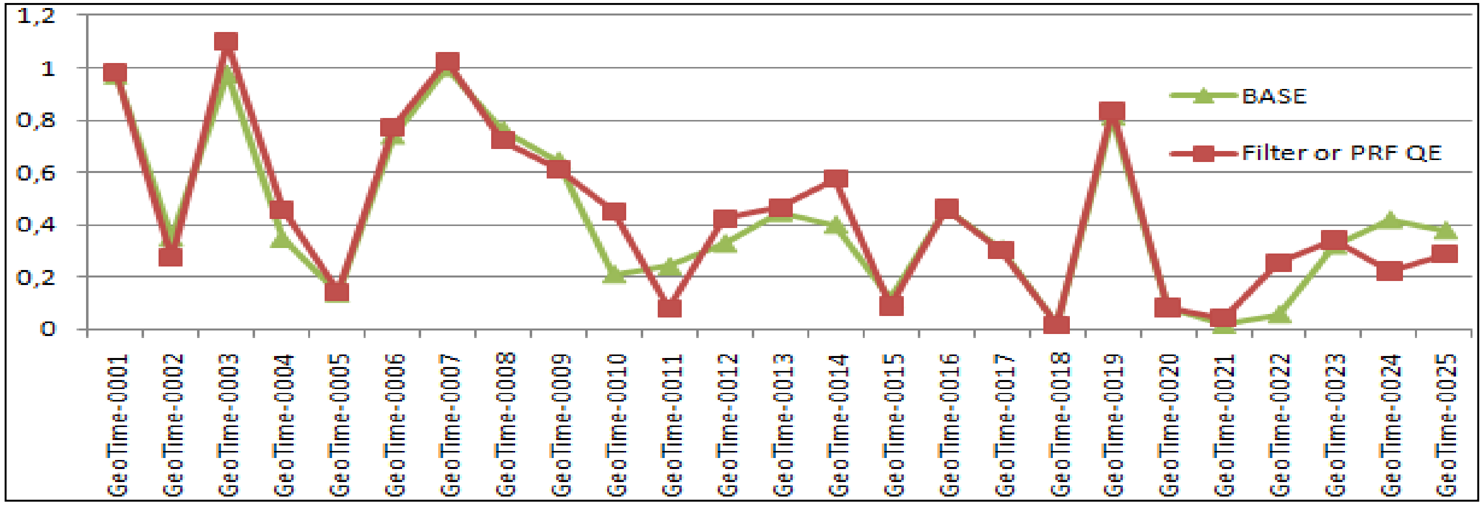
- Sentences are very fine grained …this was very restrictive excluding relevant results.

- Combination of Scores using BM25 require, **at least,** Field Score Normalization.
  - In several topics the score of the places and timexes overlapped the keywords.

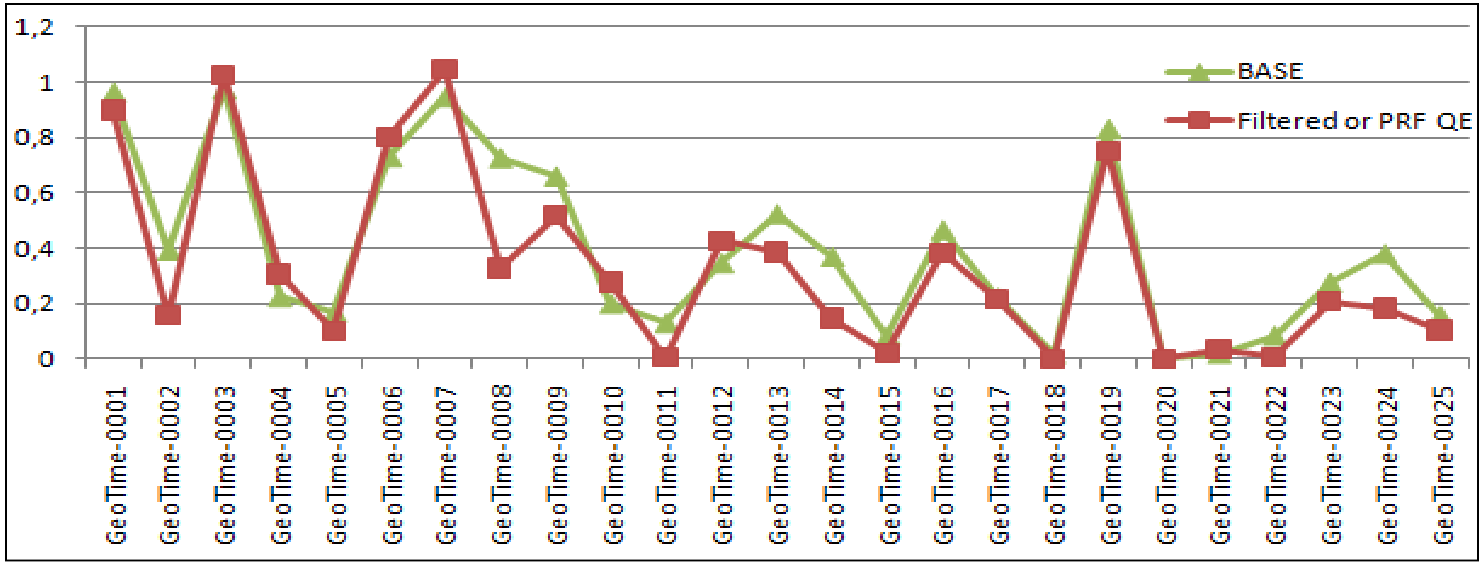
- The Combination RUN-01 (Sentences and Text) boosts the previous problems.

BASE vs Filter or PRF QE

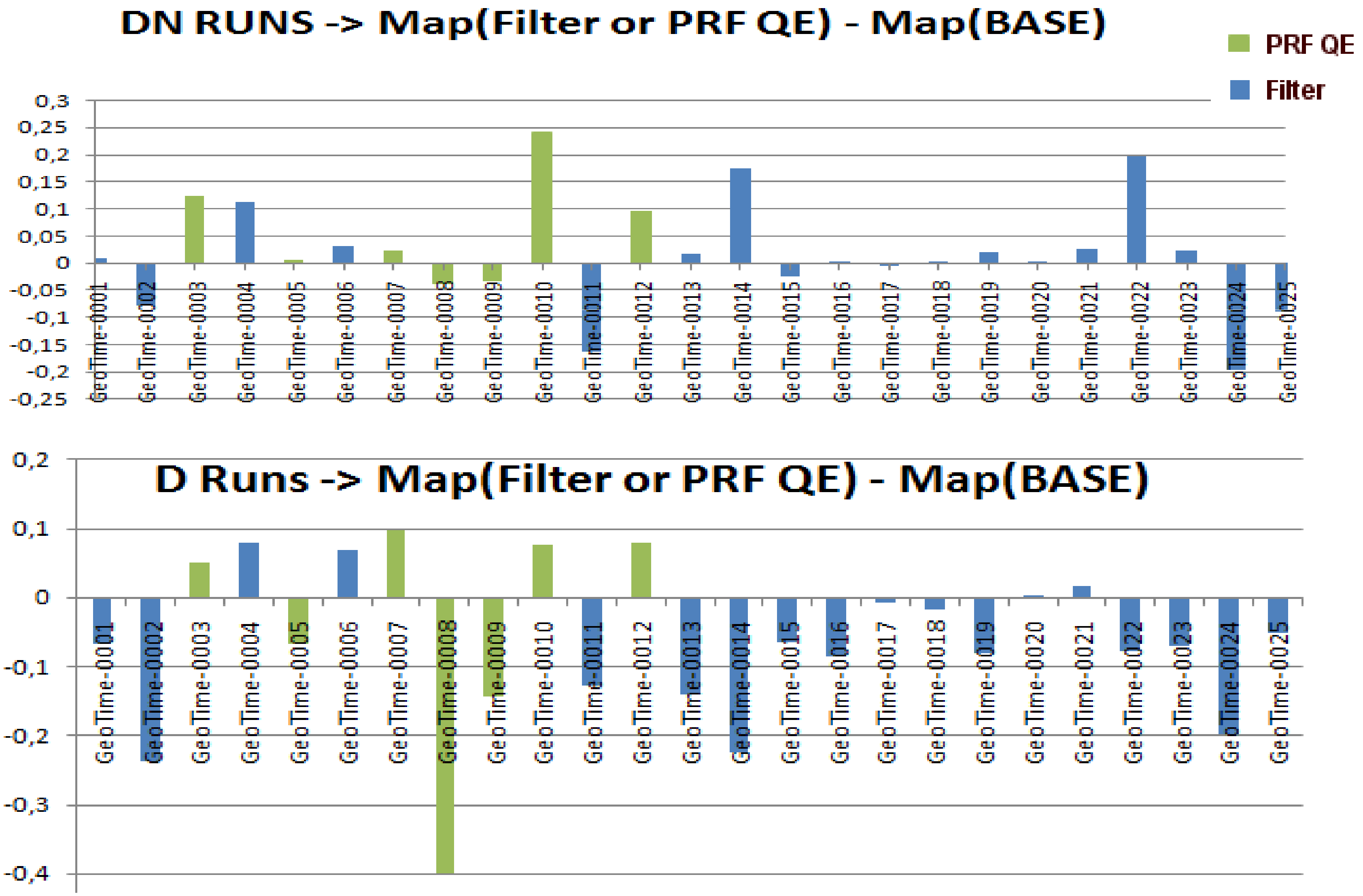
Description and Narrative “DN” RUNS



Description “D” RUNS



Filter vs Query Expansion



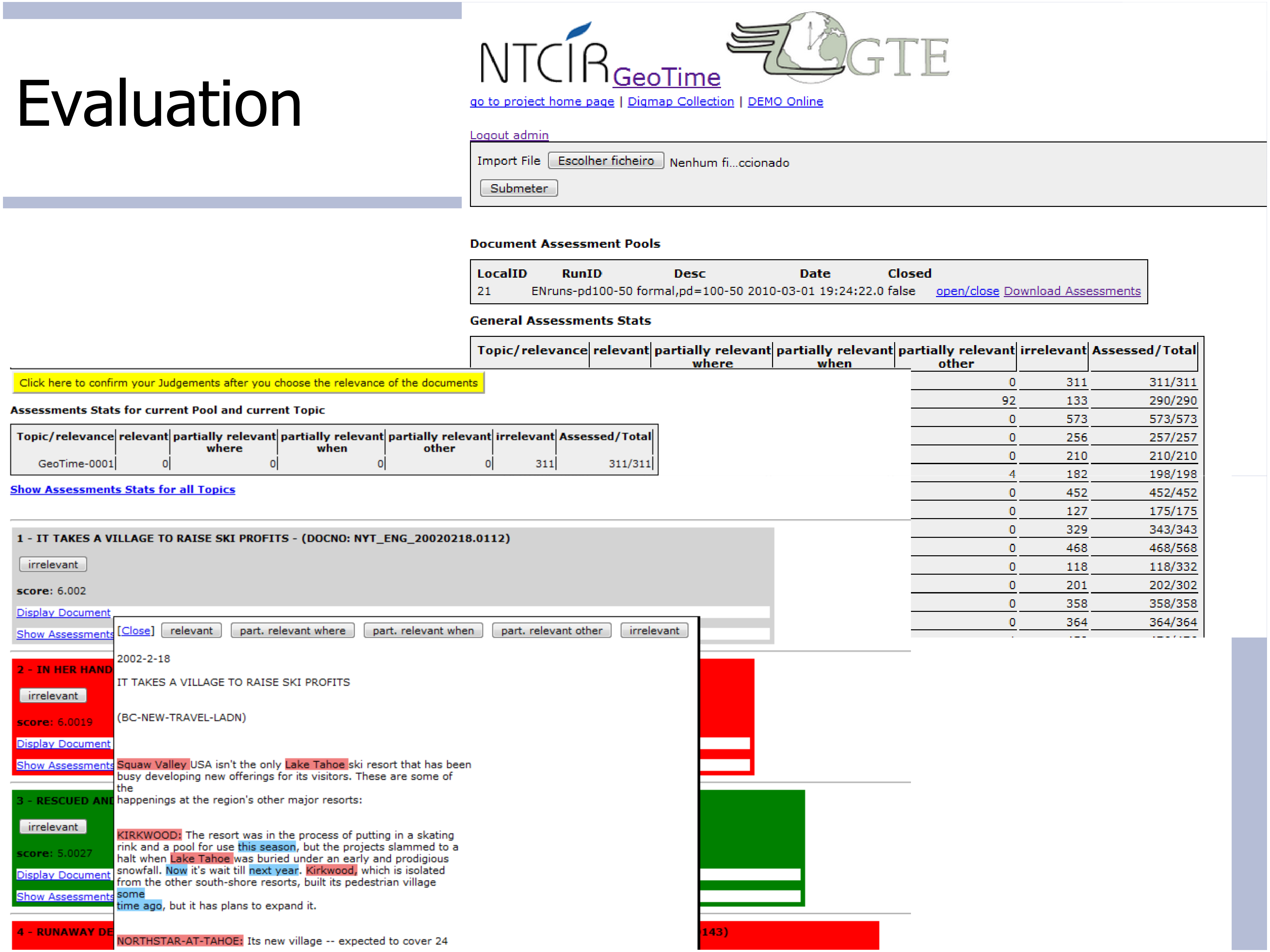
Future Challenges

- Find ways to improve Temporal Indexing to include all expressions ignored in this experiment.
- Find ways to automatically extract the topic filters.
- Try Paragraphs Context instead of sentences.
- Find ways to create an Unified Score Model or experiment Mathematical models already proposed in other areas.
- Find a standard way to make GeoTemporal Evaluation.
  - A topic containing the words Where and When can't be processed like a topic setting a place as restriction (Separation is needed)
- GeoTime Retrieval should not be only based in Topics and Documents or Answers as isolated Results but for example evaluate the Traceability/History techniques.

Thank you for your time

Questions…

Evaluation



go to project home page | Digmap Collection | DEMO Online

Logout admin

Import File [Escolher ficheiro] Nenhum fi...ccionado

[Submeter]

Document Assessment Pools

LocalID	RunID	Desc	Date	Closed		
21	ENruns-pd100-50	formal,pd=100-50	2010-03-01 19:24:22.0	false	open/close	Download Assessments

General Assessments Stats

Topic/relevance	relevant	partially relevant where	partially relevant when	partially relevant other	irrelevant	Assessed/Total
				0	311	311/311
				92	133	290/290
				0	573	573/573
				0	256	257/257
				0	210	210/210
				4	182	198/198
				0	452	452/452
				0	127	175/175
				0	329	343/343
				0	468	468/568
				0	118	118/332
				0	201	202/302
				0	358	358/358
				0	364	364/364

Click here to confirm your Judgements after you choose the relevance of the documents

Assessments Stats for current Pool and current Topic

Topic/relevance	relevant	partially relevant where	partially relevant when	partially relevant other	irrelevant	Assessed/Total
GeoTime-0001	0	0	0	0	311	311/311

Show Assessments Stats for all Topics

1 - IT TAKES A VILLAGE TO RAISE SKI PROFITS - (DOCNO: NYT_ENG_20020218.0112)

[irrelevant]

score: 6.002

Display Document

Show Assessments

2 - IN HER HAND

[irrelevant]

score: 6.0019

Display Document

Show Assessments

3 - RESCUED AN

[irrelevant]

score: 5.0027

Display Document

Show Assessments

4 - RUNAWAY DE ...**143)**

[Close] [relevant] [part. relevant where] [part. relevant when] [part. relevant other] [irrelevant]

2002-2-18

IT TAKES A VILLAGE TO RAISE SKI PROFITS

(BC-NEW-TRAVEL-LADN)

Squaw Valley USA isn't the only Lake Tahoe ski resort that has been busy developing new offerings for its visitors. These are some of the happenings at the region's other major resorts:

KIRKWOOD: The resort was in the process of putting in a skating rink and a pool for use this season, but the projects slammed to a halt when Lake Tahoe was buried under an early and prodigious snowfall. Now it's wait till next year. Kirkwood, which is isolated from the other south-shore resorts, built its pedestrian village some time ago, but it has plans to expand it.

NORTHSTAR-AT-TAHOE: Its new village -- expected to cover 24

LGTE Features for Index Fields

- Multi-Indexes Isolation at Field granularity level
  - Helps with complex index management
    - E.g. Index for temporal expressions
    - E.g. Index for keywords
- Transparent Hierarchical Indexes
  - Let developer define for example an for documents as parent and a child index for pages combining the scores of both of them using the query language transparently (paragraph:digital^0.3 text:digital^0.7)
  - Using pages as documents and indexing all the text in each page produces very big and consequently indexes which are hard to maintain.

LGTE Evaluation Framework

To create experiments …

- LGTE was the official tool used in GeoTime task of NTCIR.
- Is a simple tool based in semantic expressions highlighting to help assessors on the process of create judgments.
- Manages Pools of documents and the assessments process