The Edinburgh Geoparser

A Tool to Geoparse Text

The Edinburgh Geoparser is a language processing tool designed to detect placename references in English text and ground them against an authoritative gazetteer so that they can be plotted on a map. It was developed by researchers at the Language Technology Group at the School of Informatics at the University of Edinburgh. It has been applied in a number of research projects, for example to geo-locate literature set in Edinburgh or to geo-reference historical documents on commodity trade in the 19th century.

Click on a lat/long to centre the map there.

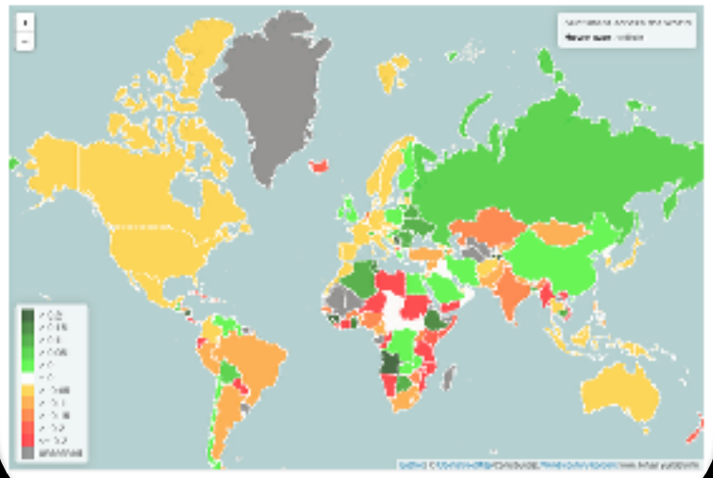
Edinburgh 55.952,-3.196 55.823,-3.093 55.950,-3.193 -37.068,

Beatrice Alex
balex@inf.ed.ac.uk, @bea_alex

THE UNIVERSITY of EDINBURGH
informatics

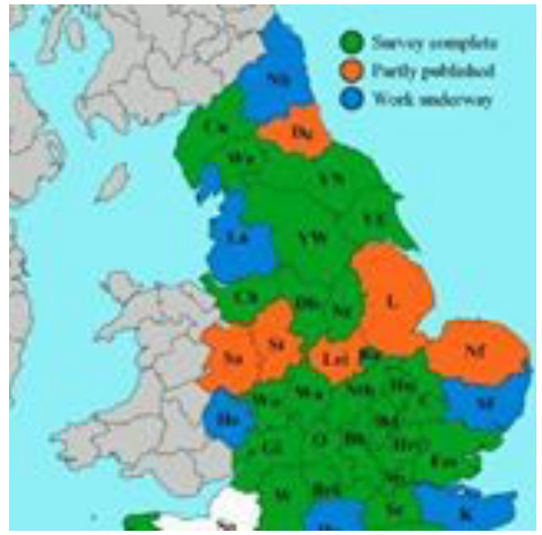
Projects

UK Connectivity

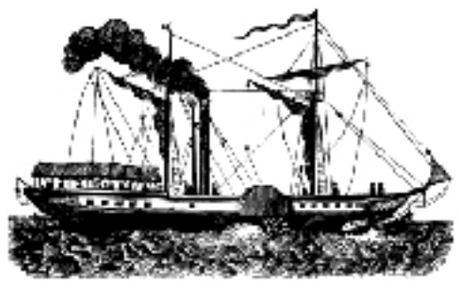


Palimpsest LitLong

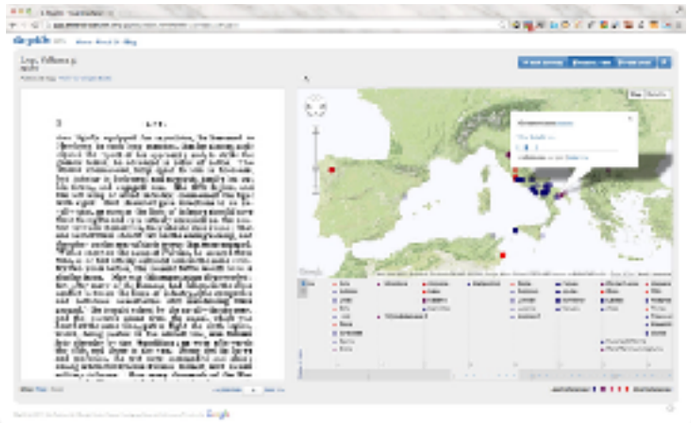
DEEP



Trading Consequences



GAP/GapVis



Historical Texts

The developers

- Claire Grover, Richard Tobin, Kate Byrne and Beatrice Alex

Goals of the workshop

- Basic overview of the Geoparser
- Geoparsing a text file
- Inspecting the visualised output
- Understanding the basic Geoparser settings
- Geoparsing multiple text files
- Extracting geo-location information from the XML
- Trying out the online demo

Software you need

- The Edinburgh Geoparser
- A terminal, e.g.
  - Terminal
  - iTerm
- A text editor (e.g. TextEdit, TextMate, TextWrangler, Komodo Edit, …)
- Web browser, e.g.
  - Safari
  - Chrome
  - Firefox

Useful material

- Edinburgh Geoparser: https://www.ltg.ed.ac.uk/software/geoparser/
- Edinburgh Geoparser Demo: http://jekyll.inf.ed.ac.uk/geoparser.html
- Documentation: http://groups.inf.ed.ac.uk/geoparser/documentation/v1.1/
- Workshop slides: http://homepages.inf.ed.ac.uk/balex/publications/geoparser-workshop.pdf

What does the Geoparser do?

- A language processing tool designed to detect place name references in English text and ground them against an authoritative gazetteer so that they can be plotted on a map.
- Works on Linux 64 bit and MacOSX.
- Is available for download free for research (University of Edinburgh GPL license)
- Contains a simple html visualiser.

Text text text text
text text text text
place name text
text text text text

Getting started

- Download the Geoparser at: https://www.ltg.ed.ac.uk/software/geoparser/
- Move the download to an appropriate location, e.g. a Software directory inside the Documents directory
- Uncompress the download (if not done automatically)
- That's it. You're ready to geoparse.

```
mv ./Downloads/geoparser-march2016.tar.gz ./Documents/Software/
cd ./Documents/software/
tar -xvf geoparser-march2016.tar
```

What's in the folder?

- Navigate to the geoparser directory and list its content:

```
cd ./geoparser-v1.1
ls
```

Name	Date Modified	Size	Kind
bin	16 Mar 2016, 10:08	--	Folder
in	16 Mar 2016, 10:08	--	Folder
lib	16 Mar 2016, 10:08	--	Folder
models	16 Mar 2016, 10:08	--	Folder
out	16 Mar 2016, 10:08	--	Folder
README	16 Mar 2016, 10:08	6 KB	TextEd...um
resolve	16 Mar 2016, 10:08	--	Folder
scripts	16 Mar 2016, 10:08	--	Folder

What's in the folder?

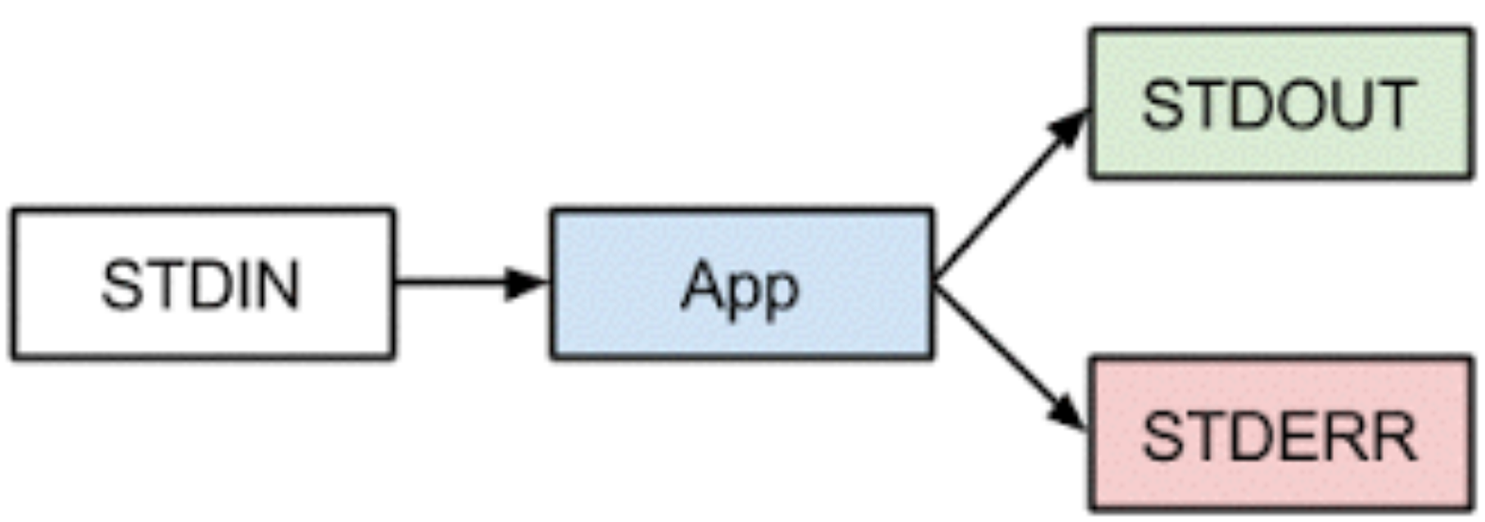
- **README**: a file with basic instructions for how to run the Geoparser
- **bin**: a set of binaries for different operating systems. We provide binaries for Linux (x86_64) and MacOSX.
- **in**: a directory with example input files
- **lib**: a set of libraries required for various processing steps
- **out**: a directory with example output files
- **resolve**: a directory containing programs required for geo-resolution
- **scripts**: a directory with a set of scripts to run the Geoparser

Basic concepts

- Command line processing
- Input/output document
- Document format
- Gazetteer
- XML

Input/Ouput

- Either stdin/sdtout (and stderr)
- Or specified input/output files



Gazetteer

gazetteer
/ˌgazəˈtɪə/

noun

a geographical index or dictionary.
"a gazetteer of place names of the Aegean"

Supported gazetteers

- More information at http://groups.inf.ed.ac.uk/geoparser/documentation/v1.0/html/gaz.html
- **GeoNames** (`-g geonames`): a world-wide gazetteer of over eight million placenames, made available free of charge.
- **OS** (`-g os`): a detailed gazetteer of UK places, derived from the Ordnance Survey 1:50,000 scale gazetteer.
- **Natural Earth** (`-g naturalearth`): a public domain vector and raster map collection of small scale (1:10m, 1:50m, 1:110m) mapping, built by the Natural Earth project.

Supported gazetteers

- **Unlock** (`-g unlock`): a comprehensive gazetteer mainly for the UK, using both OS and Natural Earth resources and augmented with major worldwide cities and countries. This is the default option on the Unlock Places service and combines all their gazetteers except DEEP.
- **DEEP** (`-g deep`): a gazetteer of historical placenames in England, built by the DEEP project (Digital Exposure of English Placenames).
- **Pleiades+** (`-g plplus`): a gazetteer of the ancient Greek and Roman world, based on the Pleiades dataset and augmented with links to Geonames.

Geoparsing a text file

- Go to the scripts directory (cd = change directory) and run the geoparser:

```
cd ./geoparser-v1.1/scripts
cat ../in/172172.txt |./run -t plain -g geonames -o ../out 172172
```

- The `cat` command prints a file to screen (or stdout). It can be fed into the geoparser as stdin.
- The `./run` command runs the geoparser. It is followed by a few parameters (-t, -g, -o) some of which need to be set to make it work.

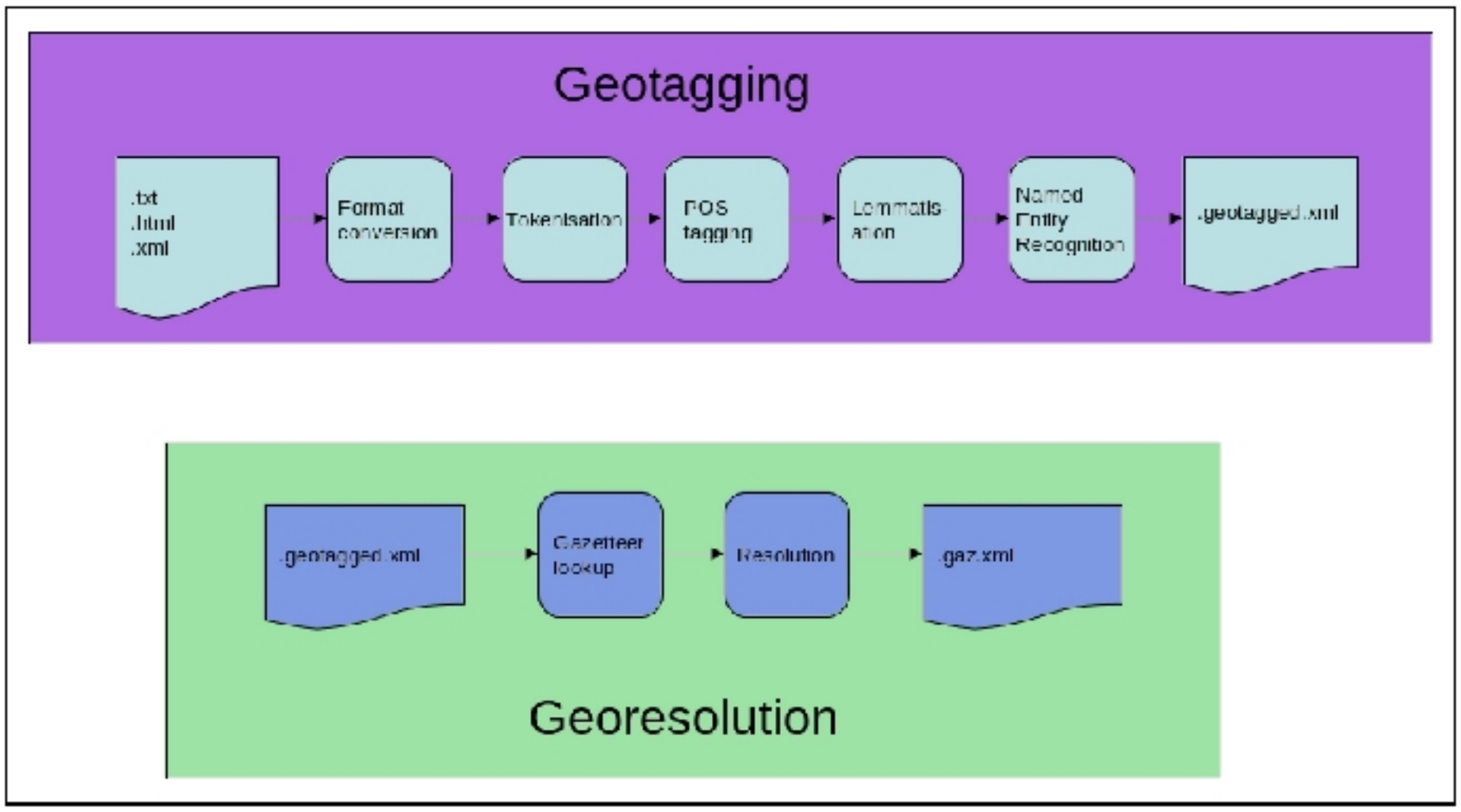
Parameters

```
./run -t plain -g geonames -o ../out 172172
```

- `-t` specifies the format of your input. (`plain`, `gb` or `ltgxml`). We recommend plain text format.
- `-g` specifies the gazetteer that should be queried. (`geonames`, `deep`, `plplus`). See gazetteer slides.
- `-o` specifies two arguments, the output directory (`../out`) and a prefix for the output file name (in this case `172172`).

What happens behind the scenes?

- The specified text file is going through a series of processing steps which are combined into one pipeline:
  - Including creation of visualisations
  - See also: http://groups.inf.ed.ac.uk/geoparser/documentation/v1.0/html/pipeline.html



Output

- To see the output files, go to the out directory:

```
cd ../out
ls 172172*
```

```
172172.display.html     172172.nertagged.xml    1721722.gazlist.html
172172.events.xml       172172.out.xml          1721722.gazmap.html
172172.gaz.xml          172172.timeline.html    1721722.geotagged.html
172172.gazlist.html     1721722.display.html    1721722.nertagged.xml
172172.gazmap.html      1721722.events.xml      1721722.out.xml
172172.geotagged.html   1721722.gaz.xml         1721722.timeline.html
```

Important output files

- `172172.out.xml`: the XML file containing the text of the file in XML including all the linguistic processing information specified inline, the named entity recognition output and the geo-resolution output specified in standoff.
- `172172.gaz.xml`: an XML file containing a ranked list of geo-resolution candidates for each extracted location mention.
- `172172.display.html`: a visual display of the geo-parsed text file containing the text, a map and a list of geo-coordinates for each extracted location.

Important output files

- **`172172.out.xml:`** the XML file containing the text of the file in XML including all the linguistic processing information specified inline, the named entity recognition output and the geo-resolution output specified in standoff.
- `172172.gaz.xml`: an XML file containing a ranked list of geo-resolution candidates for each extracted location mention.
- `172172.display.html`: a visual display of the geo-parsed text file containing the text, a map and a list of geo-coordinates for each extracted location.

Output XML file


```
<?xml version="1.0" encoding="UTF-8"?>
<document version="3">
<meta>
<attr name="docdate" id="docdate" year="2016" month="04" date="29" sdate="2016-04-29" day-number="736082"
 day="Friday" wdaynum="5">20160429</attr>
<attr name="tokeniser_version" date="20151216"/></meta>
<text>
```

```
<p><s id="s1"><w l="nadal" pws="yes" id="w16" p="NNP" group="B-NP">Nadal</w> <w pws="yes" id="w22" p="CC"
 group="O">and</w> <w wlastn="true" locname="single" l="murray" pws="yes" id="w26" p="NNP" wfirstn="true"
 group="B-NP">Murray</w> <w l="set" pws="yes" id="w33" p="VBD" headv="yes" group="B-VP">set</w> <w pws="y
es" id="w37" p="RP" group="I-VP">up</w> <w l="semus" pws="yes" id="w40" p="NN" event="true" headn="yes" g
roup="B-NP">semi</w> <w l="showdown" pws="yes" id="w45" p="NN" event="true" tmln="true" headn="yes" group
="I-NP">showdown</w>
<w pws="yes" id="w54" p="(" group="O">(</w><w l="cnn" pws="no" id="w55" p="NNP" orgname="single" group="B
-NP">CNN</w><w pws="no" id="w58" p=")" group="O">)</w> <w pws="yes" id="w60" p=":" group="O">--</w> <w wi
ki="B-wper" l="rafael" pws="yes" id="w63" p="NNP" pername="true" wfirstn="true" group="B-NP">Rafael</w> <
w wiki="I-wper" l="nadal" pws="yes" id="w70" p="NNP" group="I-NP">Nadal</w> <w pws="yes" id="w76" p="CC"
group="O">and</w> <w wiki="B-wper" l="andy" pws="yes" id="w80" p="NNP" pername="true" wfirstn="true" grou
p="B-NP">Andy</w> <w wiki="I-wper" wlastn="true" locname="single" l="murray" pws="yes" id="w85" p="NNP" w
firstn="true" group="I-NP">Murray</w> <w l="be" pws="yes" id="w92" p="VBP" headv="yes" group="B-VP">are</
w> <w pws="yes" id="w96" p="DT" headn="yes" group="B-NP">both</w> <w pws="yes" id="w101" p="IN" group="B-
PP">through</w> <w pws="yes" id="w109" p="TO" group="B-PP">to</w> <w pws="yes" id="w112" p="DT" group="B-
NP">the</w> <w l="semifinal" pws="yes" id="w116" p="NNS" event="true" headn="yes" group="I-NP">semifinals
</w> <w pws="yes" id="w127" p="IN" group="B-PP">of</w> <w pws="yes" id="w130" p="DT" group="B-NP">the</w>
 <w wlastn="true" l="roger" pws="yes" id="w134" p="NNP" wfirstn="true" headn="yes" group="I-NP">Rogers</w
> <w common="true" l="cup" pws="yes" id="w141" p="NNP" headn="yes" group="I-NP">Cup</w> <w pws="yes" id="
```

Output XML file

```
<standoff>
<ents source="ner-rb">
  <ent id="rb1" type="person">
    <parts>
      <part ew="w16" sw="w16">Nadal</part>
    </parts>
  </ent>
  <ent id="rb2" type="person">
    <parts>
      <part ew="w26" sw="w26">Murray</part>
    </parts>
  </ent>
  <ent id="rb4" type="person">
    <parts>
      <part ew="w70" sw="w63">Rafael Nadal</part>
    </parts>
  </ent>
  <ent id="rb5" type="person">
    <parts>
      <part ew="w85" sw="w80">Andy Murray</part>
    </parts>
  </ent>
  <ent id="rb6" type="location" lat="43.70011" long="-79.4163" gazref="geonames:6167865" in-country="CA" feat-type=
ppla" pop-size="4612191">
    <parts>
      <part ew="w148" sw="w148">Toronto</part>
    </parts>
  </ent>
  <ent date="22" month="05" year="2016" sdate="2016-05-22" day-number="736105" id="rb7" type="date" day="Sunday" wo
ynum="7">
    <parts>
      <part ew="w204" sw="w204">Sunday</part>
    </parts>
  </ent>
```

Important output files

- `172172.out.xml`: the XML file containing the text of the file in XML including all the linguistic processing information specified inline, the named entity recognition output and the geo-resolution output specified in standoff.
- `172172.gaz.xml`: an XML file containing a ranked list of geo-resolution candidates for each extracted location mention.
- `172172.display.html`: a visual display of the geo-parsed text file containing the text, a map and a list of geo-coordinates for each extracted location.

Important output files

- `172172.out.xml`: the XML file containing the text of the file in XML including all the linguistic processing information specified inline, the named entity recognition output and the geo-resolution output specified in standoff.
- **`172172.gaz.xml`**: an XML file containing a ranked list of geo-resolution candidates for each extracted location mention.
- `172172.display.html`: a visual display of the geo-parsed text file containing the text, a map and a list of geo-coordinates for each extracted location.

Ranked locations


```
<placenames>
 <placename id="rb6" name="Toronto">
  <place rank="1" score="1.757636992" scaled_type="0.8" scaled_pop="0.9327814568" scaled_contained_by="0" scaled_contai
ns="0" scaled_near="0" in-cc="CA" long="-79.4163" lat="43.70011" type="ppla" gazref="geonames:6167865" name="Toronto" p
op="4612191" clusteriness="891.8440693" scaled_clusteriness="0.02485553568" clusteriness_rank="10" locality="0" distanc
e-to-known="99999" scaled_known="0"/>
  <place rank="2" score="1.358629984" scaled_type="0.4" scaled_pop="0.9327814568" scaled_contained_by="0" scaled_contai
ns="0" scaled_near="0" in-cc="CA" long="-79.66632" lat="43.60012" type="rgn" gazref="geonames:6167864" name="Toronto" p
op="4612191" clusteriness="887.7750719" scaled_clusteriness="0.02584852692" clusteriness_rank="8" locality="0" distance
-to-known="99999" scaled_known="0"/>
  <place rank="3" score="1.157739277" scaled_type="0.2" scaled_pop="0.9327814568" scaled_contained_by="0" scaled_contai
ns="0" scaled_near="0" in-cc="CA" long="-79.61286" lat="43.68066" type="fac" gazref="geonames:6296338" name="Toronto Pe
arson International Airport" pop="4612191" clusteriness="891.4240778" scaled_clusteriness="0.0249578198" clusteriness_r
ank="9" locality="0" distance-to-known="99999" scaled_known="0"/>
  <place rank="4" score="0.6922152501" scaled_type="0.6" scaled_pop="0" scaled_contained_by="0" scaled_contains="0" sca
led_near="0" in-cc="US" long="-92.52546" lat="38.00365" type="ppl" gazref="geonames:4411872" name="Toronto" clusterines
s="653.9875787" scaled_clusteriness="0.09221525012" clusteriness_rank="1" locality="0" distance-to-known="99999" scaled
_known="0"/>
  <place rank="5" score="0.6842565066" scaled_type="0.6" scaled_pop="0" scaled_contained_by="0" scaled_contains="0" sca
led_near="0" in-cc="US" long="-89.62982" lat="39.71394" type="ppl" gazref="geonames:4251360" name="Toronto" clusterines
s="678.4017926" scaled_clusteriness="0.08425650659" clusteriness_rank="2" locality="0" distance-to-known="99999" scaled
_known="0"/>
  <place rank="6" score="0.6837386385" scaled_type="0.6" scaled_pop="0" scaled_contained_by="0" scaled_contains="0" sca
led_near="0" in-cc="US" long="-87.49557" lat="39.7817" type="ppl" gazref="geonames:4265888" name="Toronto" clusteriness
="680.021624" scaled_clusteriness="0.08373863848" clusteriness_rank="3" locality="0" distance-to-known="99999" scaled_k
nown="0"/>
  <place rank="7" score="0.6767765887" scaled_type="0.6" scaled_pop="0" scaled_contained_by="0" scaled_contains="0" sca
led_near="0" in-cc="US" long="-96.64255" lat="44.57302" type="ppl" gazref="geonames:5232379" name="Toronto" clusterines
s="702.1773599" scaled_clusteriness="0.07677658872" clusteriness_rank="4" locality="0" distance-to-known="99999" scaled
_known="0"/>
  <place rank="8" score="0.6603991986" scaled_type="0.6" scaled_pop="0" scaled_contained_by="0" scaled_contains="0" sca
led_near="0" in-cc="US" long="-90.86403" lat="41.90502" type="ppl" gazref="geonames:4878739" name="Toronto" clusterines
s="757.1843043" scaled_clusteriness="0.06039919856" clusteriness_rank="5" locality="0" distance-to-known="99999" scaled
_known="0"/>
```

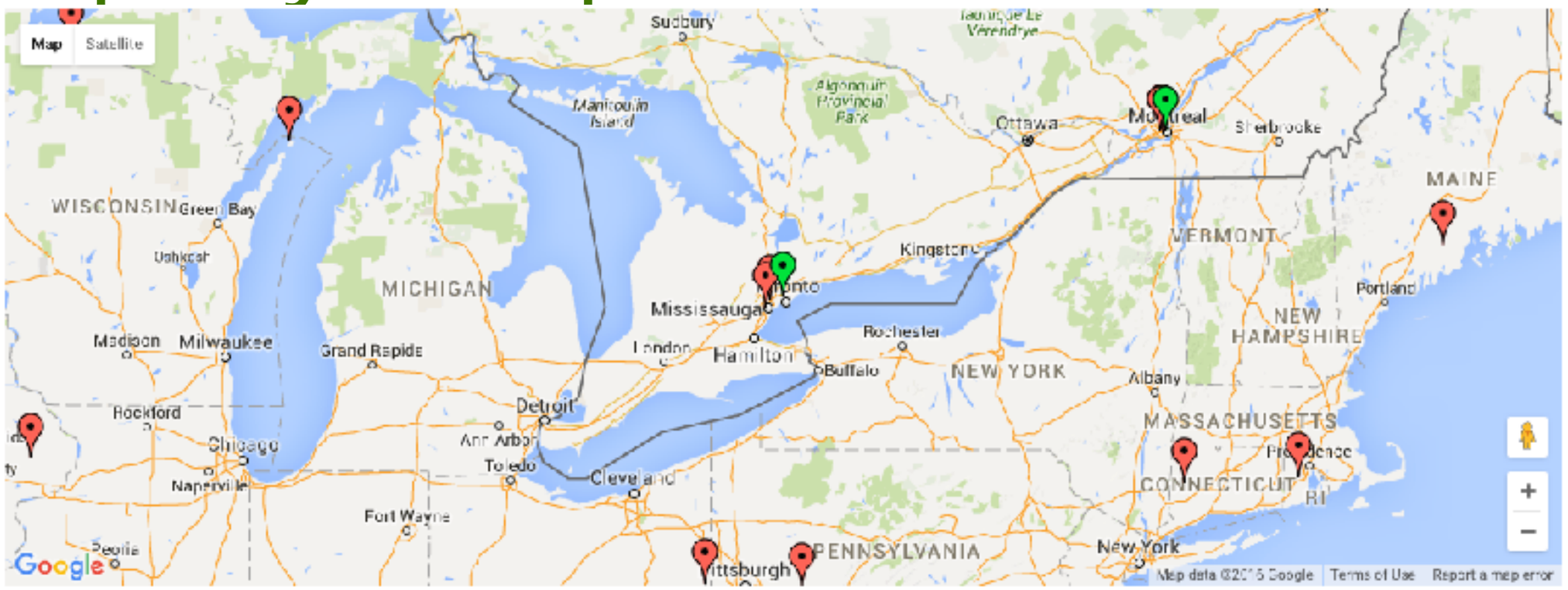
Important output files

- `172172.out.xml`: the XML file containing the text of the file in XML including all the linguistic processing information specified inline, the named entity recognition output and the geo-resolution output specified in standoff.
- `172172.gaz.xml`: an XML file containing a ranked list of geo-resolution candidates for each extracted location mention.
- `172172.display.html`: a visual display of the geo-parsed text file containing the text, a map and a list of geo-coordinates for each extracted location.

Important output files

- `172172.out.xml`: the XML file containing the text of the file in XML including all the linguistic processing information specified inline, the named entity recognition output and the geo-resolution output specified in standoff.
- `172172.gaz.xml`: an XML file containing a ranked list of geo-resolution candidates for each extracted location mention.
- **`172172.display.html`**: a visual display of the geo-parsed text file containing the text, a map and a list of geo-coordinates for each extracted location.

Display output in a browser



Nadal and Murray set up semi showdown (CNN) -- Rafael Nadal and Andy Murray are both through to the semifinals of the Rogers Cup in Toronto, where they will face each other for a place in Sunday's final. Murray played some superb tennis in crushing the in-form David Nalbandian but Nadal had to recover from dropping the opening set to get past Germany's Philipp Kohlschreiber. Nalbandian won the ATP title in Washington last weekend and came into Friday's encounter on an 11-match unbeaten streak. But the fourth-seeded Briton, who has struggled to find his top form this season, brushed aside the Argentine in impressive fashion, securing a 6-2 6-2 victory in just 69 minutes. "It was probably one of the best matches I've played this year," Murray told the official ATP Tour website. "I served well and got the first good hits in a lot of the rallies, so I was able to dictate a lot of the points," added Murray -- who is the defending champion after winning the tournament in Montreal last year. Meanwhile, top seed Nadal also secured his place in the last four, but he was not as impressive as Murray in a 3-6 6-3 6-4 victory over Kohlschreiber In the evening session, third seed Federer will face a Wimbledon re-match with Czech Tomas Berdych, who beat him in the quarterfinals of the grasscourt tournament. The winner of that match will face either second seed Novak Djokovic or Jeremy Chardy of France for a place in the final.

Click on a lat/long to centre the map there.

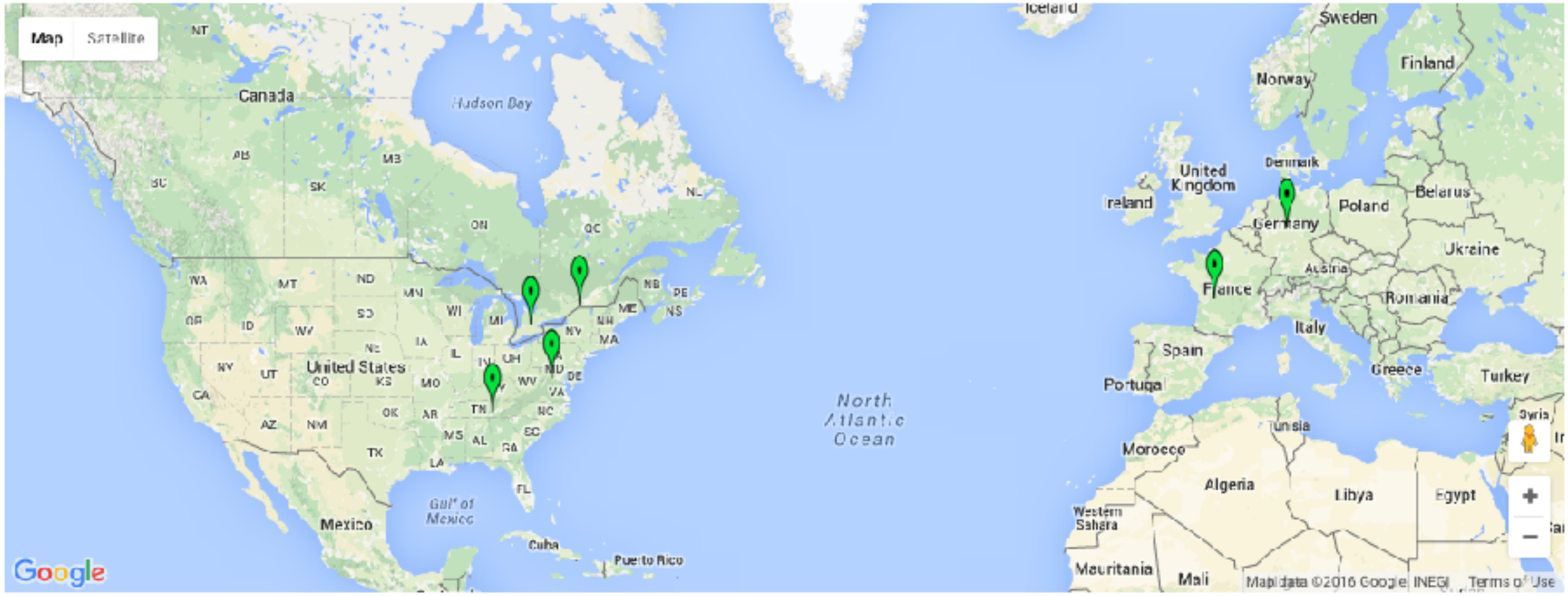
Toronto	43.700,-79.416	43.600,-79.666	43.681,-79.613	38.004,-92.525	39.
Germany	51.500,10.500	38.462,-85.543	39.003,-82.905	34.896,-83.467	40.
Washington	38.895,-77.036	54.900,-1.517	38.558,-91.012	38.659,-87.173	42.
Montreal	45.509,-73.588	45.500,-73.682	45.527,-73.653	45.500,-73.666	45.
Wimbledon	35.718,-83.979	47.170,-98.460	39.509,-76.407	35.751,-78.776	51.
France	46.000,2.000	41.649,-7.471	-21.684,34.703	43.972,-111.275	40.

Only top-ranked locations

```
cd ../scripts
cat ../in/172172.txt | ./run -t plain -g
geonames -top -o ../out 172172
```

- This creates some additional output files, most notably **`172172.display-top.html`** which only contains the top-ranked location candidates, so only the green geo-coordinate pairs and pins.

Only top-ranked locations



Nadal and Murray set up semi showdown (CNN) -- Rafael Nadal and Andy Murray are both through to the semifinals of the Rogers Cup in Toronto, where they will face each other for a place in Sunday's final. Murray played some superb tennis in crushing the in-form David Nalbandian but Nadal had to recover from dropping the opening set to get past Germany's Philipp Kohlschreiber. Nalbandian won the ATP title in Washington last weekend and came into Friday's encounter on an 11-match unbeaten streak. But the fourth-seeded Briton, who has struggled to find his top form this season, brushed aside the Argentine in impressive fashion, securing a 6-2 6-2 victory in just 69 minutes. "It was probably one of the best matches I've played this year," Murray told the official ATP Tour website. "I served well and got the first good hits in a lot of the rallies, so I was able to dictate a lot of the points," added Murray -- who is the defending champion after winning the tournament in Montreal last year. Meanwhile, top seed Nadal also secured his place in the last four, but he was not as impressive as Murray in a 3-6 6-3 6-4 victory over Kohlschreiber In the evening session, third seed Federer will face a Wimbledon re-match with Czech Tomas Berdych, who beat him in the quarterfinals of the grasscourt tournament. The winner of that match will face either second seed Novak Djokovic or Jeremy Chardy of France for a place in the final.

Click on a lat/long to centre the map there.

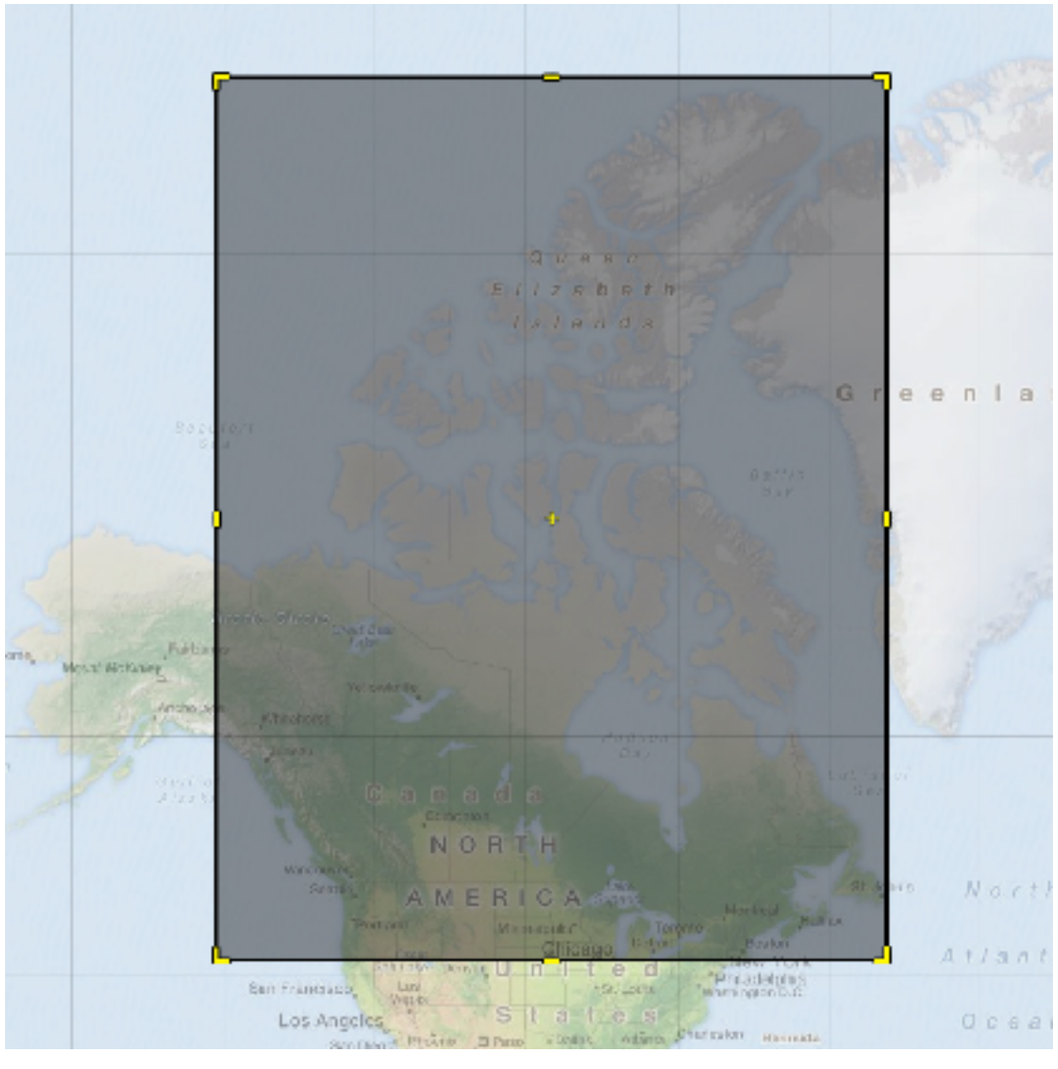
Toronto	43.700,-79.416
Germany	51.500,10.500
Washington	38.895,-77.036
Montreal	45.509,-73.588
Wimbledon	35.718,-83.979
France	46.000,2.000

Try your own example file

- If you brought a file, try geoparsing it now.
- Put the file into the `./in` directory and run the geoparser using it as the input file instead of the file in the previous example.
- If not, then try another .txt file in the `./in` directory.
- Try to change the gazetteer and see how it affects things.
- Don’t forget to change the output prefix or the output of the previous example will be overwritten!

Geographical preference

- You can specify a bounding circle (`-l locality`) or a bounding box (`-lb locality box`). The Geoparser will prefer places within the specified area but will still consider locations outside it.



Circular locality

- To specify a circular locality use the following command: `-l lat long radius score`
  - `lat` and `long` are in decimal degrees (i.e. 57.5 for 57 degrees 30 mins)
  - `radius` is specified in km
  - `score` is a numeric weight assigned to locations within the area (else 0).

Bounding box

- To specify a locality box use: `-lb W N E S score`
  - `W`(est) `N`(orth) `E`(ast) `S`(outh) are decimal degrees
  - `score` is the same as for option -l.

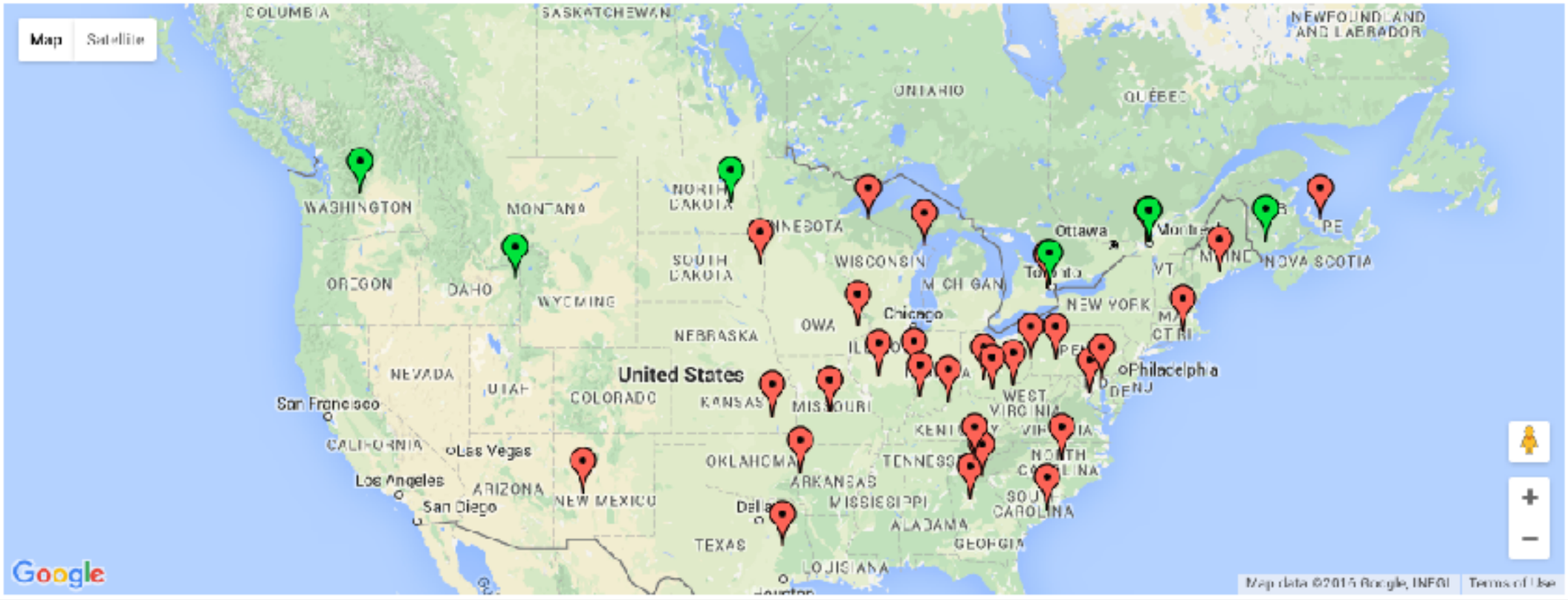
Example

- For example, a bounding box for Canada is [W:-141.002701, N:83.110619, E:-52.620201, S: 41.681019].
- Go back to the scripts directory and run the following command:

```
cat ../in/172172.txt | ./run -t plain -g geonames
-lb -141.002701 83.110619 -52.620201 41.681019 2 -
o ../out 172172
```

- The score was set to 2. This gives a location within the bounding box twice as much weight during geo-resolution as for example the population size of a location.

Example output

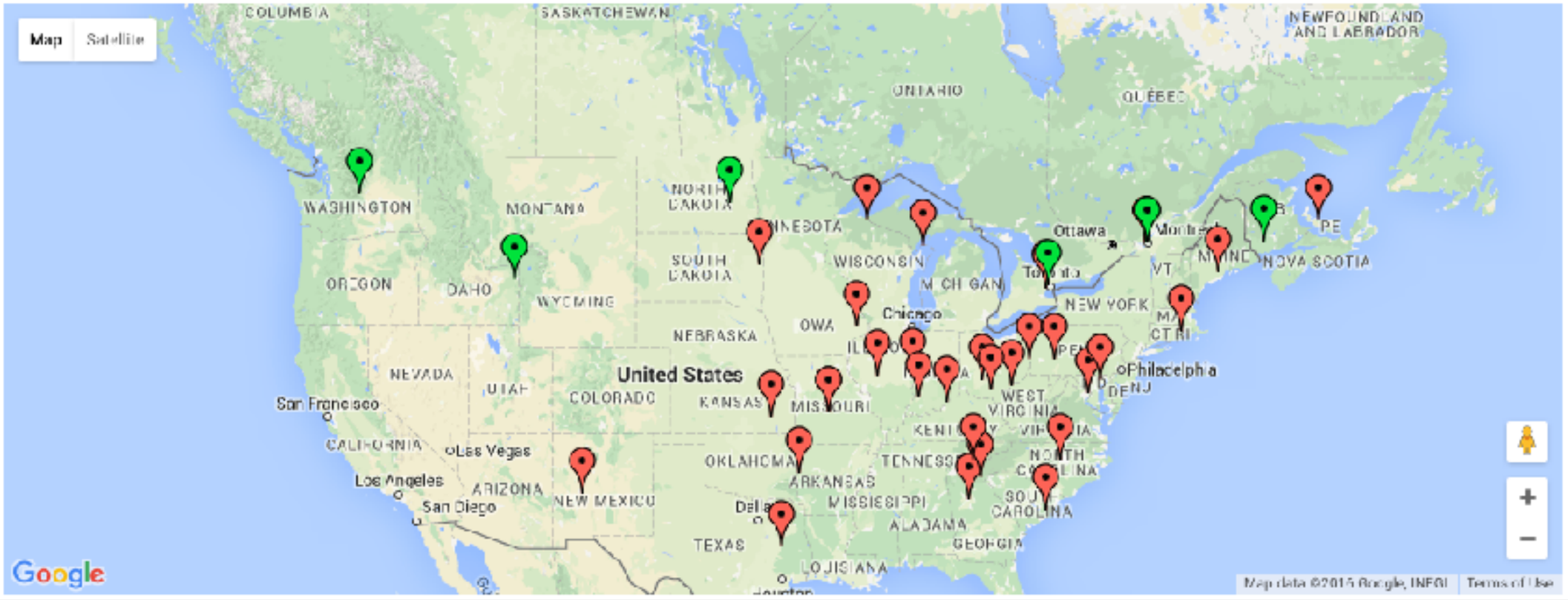


Nadal and Murray set up semi showdown (CNN) -- Rafael Nadal and Andy Murray are both through to the semifinals of the Rogers Cup in Toronto, where they will face each other for a place in Sunday's final. Murray played some superb tennis in crushing the in-form David Nalbandian but Nadal had to recover from dropping the opening set to get past Germany's Philipp Kohlschreiber. Nalbandian won the ATP title in Washington last weekend and came into Friday's encounter on an 11-match unbeaten streak. But the fourth-seeded Briton, who has struggled to find his top form this season, brushed aside the Argentine in impressive fashion, securing a 6-2 6-2 victory in just 69 minutes. "It was probably one of the best matches I've played this year," Murray told the official ATP Tour website. "I served well and got the first good hits in a lot of the rallies, so I was able to dictate a lot of the points," added Murray -- who is the defending champion after winning the tournament in Montreal last year. Meanwhile, top seed Nadal also secured his place in the last four, but he was not as impressive as Murray in a 3-6 6-3 6-4 victory over Kohlschreiber. In the evening session, third seed Federer will face a Wimbledon re-match with Czech Tomas Berdych, who beat him in the quarterfinals of the grasscourt tournament. The winner of that match will face either second seed Novak Djokovic or Jeremy Chardy of France for a place in the final.

Click on a lat/long to centre the map there.

Toronto	43.700,-79.416	43.600,-79.666	43.681,-79.613	44.573,-96.643	41
Germany	45.567,-66.632	51.500,10.500	38.462,-85.543	39.003,-82.905	34
Washington	47.500,-120.501	45.395,-86.932	44.274,-69.367	41.688,-71.567	38
Montreal	45.509,-73.588	45.500,-73.682	45.527,-73.653	45.500,-73.666	45
Wimbledon	47.170,-98.460	35.718,-83.979	39.509,-76.407	35.751,-78.776	51
France	43.972,-111.275	46.000,2.000	41.649,-7.471	-21.684,34.703	40

Example output



- **Note:** The locality option should be used with care and should ideally only be applied for documents where you are relatively certain that all or the majority of locations appear within the specified area.

year. Meanwhile, top seed Nadal also secured his place in the last four, but he was not as impressive as Murray in a 3-6 6-3 6-4 victory over Kohlschreiber In the evening session, third seed Federer will face a Wimbledon re-match with Czech Tomas Berdych, who beat him in the quarterfinals of the grasscourt tournament. The winner of that match will face either second seed Novak Djokovic or Jeremy Chardy of France for a place in the final.

Specifying a document date

- `-d` specifies the document date (`YEAR-MONTH-DATE`). This parameter is optional. It is used for normalisation of temporal expressions in the document, for example to be able to compute which particular date the string “Sunday” refers to.

```
cat ../in/172172.txt | ./run -t plain -g
geonames -d 2010-08-10 -o ../out 172172
```

- Date information stored:
  - `sdate` refers to the grounded date expressed as a string,
  - `day-number` refers to a unique day number where 1 corresponds to the 1st of January 1 AD, and
  - `wdaynum` refers to the week day number where 1 corresponds to Monday, 2 to Tuesday etc.

Output example

```
<?xml version="1.0" encoding="UTF-8"?>
<document version="3">
<meta>
<attr name="docdate" id="docdate" year="2010" month="08" date="10" sdate="2010-08-10" day-n
umber="733993" day="Tuesday" wdaynum="2">20100810</attr>
<attr name="tokeniser_version" date="20151216"/></meta>
<text>



<p><s id="s1"><w l="nadal" pws="yes" id="w16" p="NNP" group="B-NP">Nadal</w> <w pws="yes" i
d="w22" p="CC" group="O">and</w> <w wlastn="true" locname="single" l="murray" pws="yes" id=
"w26" p="NNP" wfirstn="true" group="B-NP">Murray</w> <w l="set" pws="yes" id="w33" p="VBD"
headv="yes" group="B-VP">set</w> <w pws="yes" id="w37" p="RP" group="I-VP">up</w> <w l="sem
us" pws="yes" id="w40" p="NN" event="true" headn="yes" group="B-NP">semi</w> <w l="showdown
" pws="yes" id="w45" p="NN" event="true" tmln="true" headn="yes" group="I-NP">showdown</w>

...

 <ent date="15" month="08" year="2010" sdate="2010-08-15" day-number="733998" id="rb7"
type="date" day="Sunday" wdaynum="7">
   <parts>
     <part ew="w204" sw="w204">Sunday</part>
   </parts>
 </ent>
```

Geo-parsing multiple files

- You can download a script to do that here: http://groups.inf.ed.ac.uk/geoparser/scripts/run-multiple-files.sh
- Move it to the scripts directory and make it executable.

```
chmod u+x run-multiple-files.sh
```

- Let's open it to see what it does.

```
#!/bin/sh
# Author: Beatrice Alex
# Date: 28-01-2016
# Description: Run the Geoparser on multiple text files

usage="./run-multiple-files -i inputDirectory -o outputDirectory"

# check that some parameters are specified. If not is specified, then
# the script is exited and the usage is printed
if [ $# -eq "0" ]
then
    echo "No arguments specified"
    echo "usage: $usage" >&2
    exit 2
fi

# while loop to set up the arguments specified when running the script. If the
# arguments are wrong it exits the script and prints the usage
while test $# -gt 0
do
    arg=$1
    shift
    case $arg in
    -i)
        inputdirname=$1
        shift 1
        ;;
    -o)
        outputdirname=$1
        shift 1
        ;;
    *)
        echo "Wrong argument specified"
        echo "usage: $usage" >&2
        exit 2
    esac
done

# for loop to list a set of text files specified in the input directory
for i in `ls $inputdirname/[1br]*.txt`
do
    # a print statement to say which file is currently being processed
    echo Processing $i

    # the prefix is derived from the file name, i.e. everything before the format
    # extension ".txt"
    prefix=`basename $i ".txt"`

    # each file is then geo-parsed and the output is written to the output directory
    cat $i | ./run -t plain -g geonames -o $outputdirname $prefix
done
```

Geo-parsing multiple files

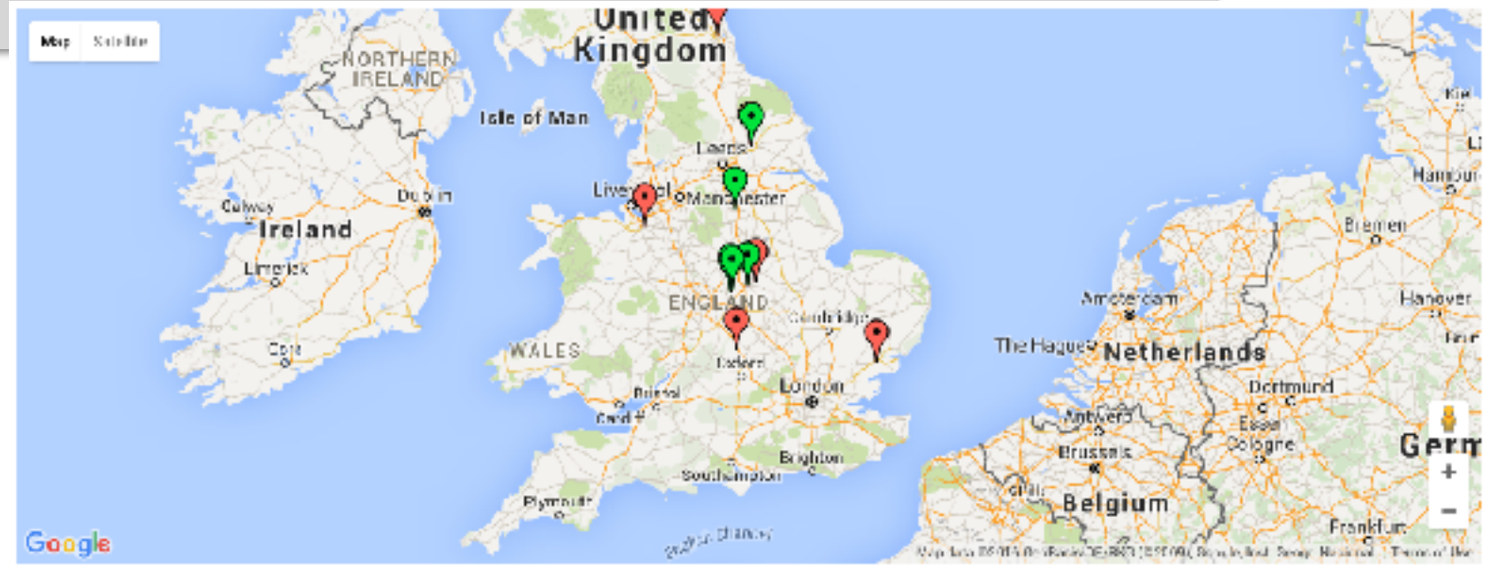
```
./run-multiple-files.sh -i ../in -o ../out
```

- `-i` specifies the input directory and
- `-o` the output directory
- When you run this command in your scripts directory you should see:

```
Processing ../in/172172.txt
Processing ../in/burtons.txt
Processing ../in/richard111.txt
```

Geo-parsing multiple files

```
open ../out/richard111.display.html (on MacOS)
xdg-open ../out/richard111.display.html (on Linux)
```



Leicester prepares for Richard III reburial

Dadlington usually divides its sympathies between the two sides, but for this they are coming out in favour of the deposed king in the Battle of Bosworth Field

Leicester is filling up with yellow road signs, warning of the return of the once and future king: "King Richard III Route Sunday 22 – expect delays".

Many have been wondering how to get hold of them to keep as souvenirs of an extraordinary episode in the history of the city and county, as the final preparations get underway for the 21st century reburial of a medieval king.

It is more than two years since the bones rediscovered by archaeologists under a council car park in the heart of Leicester, were formally identified as those of the last Plantagenet king.

In 1485 the burial by Franciscan friars was unceremonious. In the heat of August, several days after his death at Bosworth, he was buried hastily in a roughly-dug grave slightly too small for his naked body. This time a week of events will surround the reinterment on Thursday 26 in the cathedral – delayed for a year by a legal challenge claiming that he should have been buried in York

Click on a lat/long to centre the map there.

Leicester	52.639,-1.132	52.633,-1.133	52.667,-1.000	42.772,-77.8
Dadlington	52.579,-1.405			
Bosworth Field				
Bosworth	43.783,-80.650	33.023,-85.561	39.470,-93.335	-20.200,146.
York	53.958,-1.083	53.961,-1.091	43.696,-79.451	38.905,-82.1
Vaughns				
Red Lion	53.363,-1.333	52.017,-1.310	39.901,-76.606	40.481,-75.6
Market Bosworth	52.624,-1.402	52.622,-1.415		
Sutton Cheney	52.601,-1.385	52.593,-1.411		
St James	39.563,-77.758	40.879,-73.157	-32.000,115.910	10.672,-61.5
Ohio	40.250,-83.000	45.322,-74.978	32.583,-83.816	41.558,-89.4
Ambion Hill				

- Rerun the geoparser just on this file but specifying the gazetteer "`deep`" (English place names).

Geo-parsing multiple files

- Geo-parsing multiple files on the command line (without a script):

```
for i in `ls ../in/[1br]*.txt`; do echo Processing
$i; prefix=`basename $i ".txt"`; cat $i | ./run -t
plain -g geonames -o ../out $prefix ; done
```

Extracting geo-information to TSV

- The aim is to extract all location entities in a Geoparser XML output file and present them in TSV format.

XML > TSV

LX-XML2

- The Geoparser is distributed with a useful set of XML processing tools called `LT-XML2` authored by Richard Tobin. The binaries for these tools are located in the `./geoparser-v1.1/bin` directory, inside:
  - `sys-i386-64`: if you using a 64 bit Linux machine or
  - `sys-i386-snow-leopard`: if you're using MacOSX. These should work for all MacOSX installations and not just on Snow Leopard.
- These tools work in combination with XPATH expressions. The best tool for printing out information stored in XML is `lxprintf`.

Data extraction example

```
  <ent id="rb6" type="location" lat="43.70011" long="-79.4163" gazref="geonames:6167865" in-country="CA" feat-type=
ppla" pop-size="4612191">
    <parts>
      <part ew="w148" sw="w148">Toronto</part>
    </parts>
  </ent>
```

- On Linux use:

```
./bin/sys-i386-64/lxprintf -e "ent[@type='location']"
"%s\t%s\t%s\t%s\t%s\n" "normalize-space(parts)" "@gazref"
"@in-country" "@lat" "@long" < ./out/burtons.out.xml > ./
out/burtons.out.tsv
```

- and on MacOSX type:

```
./bin/sys-i386-snow-leopard/lxprintf -e
"ent[@type='location']" "%s\t%s\t%s\t%s\t%s\n" "normalize-
space(parts)" "@gazref" "@in-country" "@lat" "@long" < ./
out/burtons.out.xml > ./out/burtons.out.tsv
```

lxprintf

```
  <ent id="rb6" type="location" lat="43.70011" long="-79.4163" gazref="geonames:6167865" in-country="CA" feat-type=
ppla" pop-size="4612191">
    <parts>
      <part ew="w148" sw="w148">Toronto</part>
    </parts>
  </ent>
```

- `normalize-space(parts)` refers to the location mention recognised in the text.
- `@gazref` refers to the ID reference of the location in the gazetteer, if resolved.
- `@in-country` refers to the country the location appears in, if resolved.
- `@lat` refers to the latitude of the location, if resolved.
- `@long` refers to the longitude of the location, if resolved.
- `@feat-type` refers to the feature type of the location.

TSV Output

- The resulting TSV output contains the location name, the GeoNames identifier, the country the location is in, and the latitude and longitude coordinates.

```
sunny:geoparser-v1.1 balex$ cat ./out/burtons.out.tsv
Wirral  geonames:7733088        GB      53.37616        -3.10501
Moreton geonames:2642204        GB      53.4    -3.11667
Moreton geonames:2642204        GB      53.4    -3.11667
Wirral borough  geonames:7733088        GB      53.37616        -3.10501
Wirral  geonames:7733088        GB      53.37616        -3.10501
Moreton geonames:2642204        GB      53.4    -3.11667
Moreton geonames:2642204        GB      53.4    -3.11667
```

Extraction script

- Download it here: http://groups.inf.ed.ac.uk/geoparser/scripts/extract-to-tsv.sh
- Move the file to the scripts directory and make it executable:

```
chmod u+x extract-to-tsv.sh
```

- Then run

```
./extract-to-tsv.sh < ../out/burtons.out.xml
> ../out/burtons.out.tsv
```

- Try to change the script to print out the feature type of the location in an additional column.

Using the online demo

- Demo: http://jekyll.inf.ed.ac.uk/geoparser.html
- It is only a visual interface to the Geoparser and its output.
- It allows you to upload a text file and select a gazetteer. We recommend selecting the GeoNames gazetteer as it has global coverage. You can use one of the text files provided in the Geoparser distribution (e.g. `geoparser/in/172172.txt`) to try it out.
- **Note:** The demo is not as configurable as the download and should only be used to try out small examples and not for geoparsing a large number of files.

Using the online demo

Upload a plain TXT or XML file: Choose File no file selected
Choose a gazetteer: geonames ?
Go!

type is text/plain (plain), gazetteer is geonames

Please wait

Done

Nadal and Murray set up semi showdown (CNN) -- Rafael Nadal and Andy Murray are both through to the semifinals of the Rogers Cup in Toronto, where they will face each other for a place in Sunday's final. Murray played some superb tennis in crushing the in-form David Nalbandian but Nadal had to recover from dropping the opening set to get past Germany's Philipp Kohlschreiber. Nalbandian won the ATP title in Washington last weekend and came into Friday's encounter on an 11-match unbeaten streak. But the fourth-seeded Briton, who has struggled to find his top form this season, brushed aside the Argentine in impressive fashion, securing a 6-2 6-2 victory in just 69 minutes. "It was probably one of the best matches I've played this year," Murray told the official ATP Tour website. "I served well and got the first good hits in a lot of the rallies, so I was able to dictate a lot of the points," added Murray -- who is the defending champion after winning the tournament in Montreal last year. Meanwhile, top seed Nadal also secured his place in the last four, but he was not as impressive as Murray in a 3-6 6-3 6-4 victory over Kohlschreiber In the evening session, third seed Federer will face a Wimbledon re-match with Czech Tomas Berdych, who beat him in the quarterfinals of the grasscourt tournament. The winner of that match will face either second seed Novak Djokovic or Jeremy Chardy of France for a place in the final.

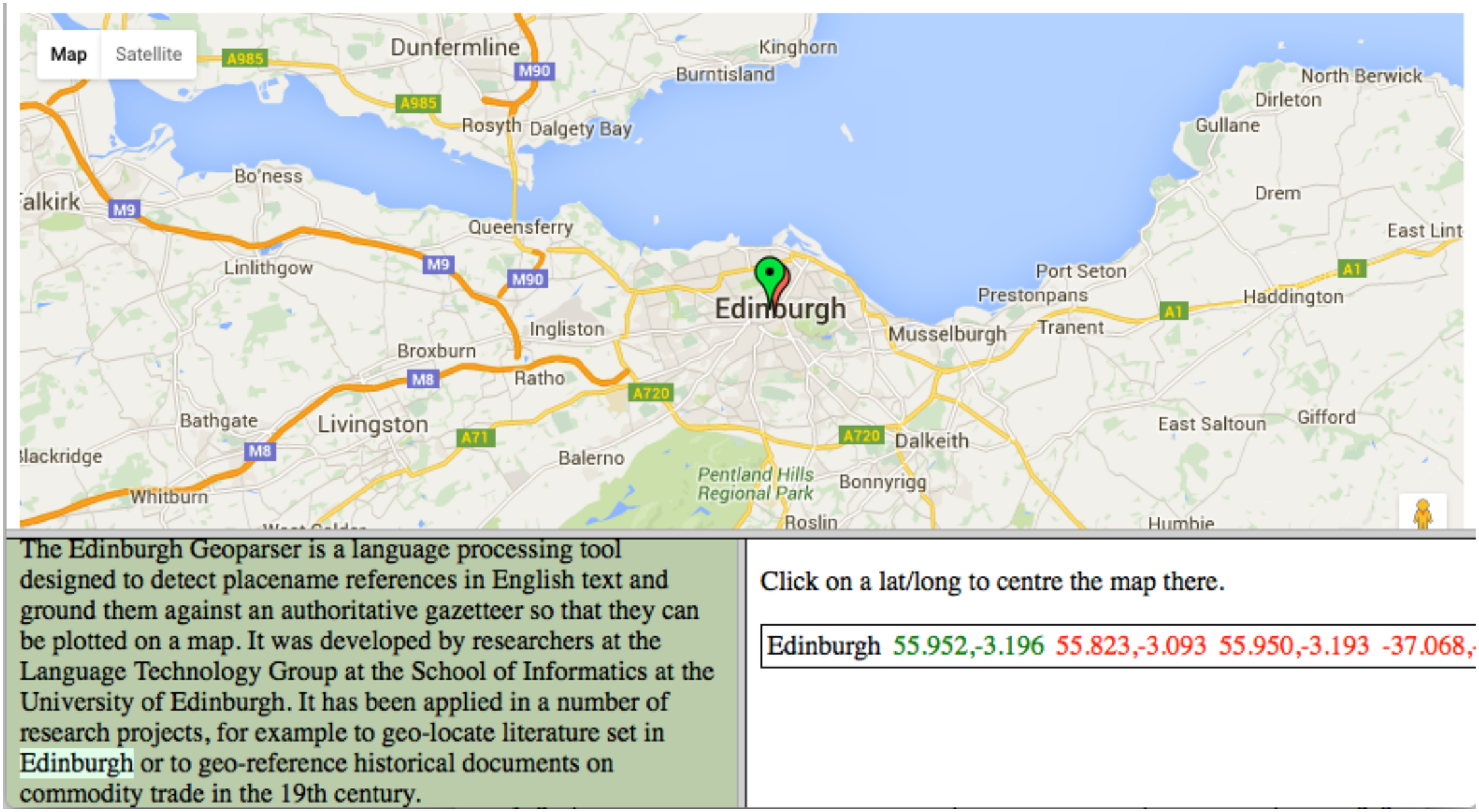
Click on a lat/long to centre the map there.

Toronto	43.700,-79.416	43.600,-79.666	43.681,-79.613	39.782,-87.496	3
Germany	51.500,10.500	38.462,-85.543	39.003,-82.905	34.896,-83.467	4
Washington	38.895,-77.036	47.500,-120.501	54.900,-1.517	38.659,-87.173	3
Montreal	45.509,-73.588	45.500,-73.682	45.527,-73.653	45.500,-73.666	4
Wimbledon	35.718,-83.979	39.509,-76.407	47.170,-98.460	35.751,-78.776	5
France	46.000,2.000	41.649,-7.471	-21.684,34.703	43.972,-111.275	4

References …

- If you use the Geoparser for your research, please cite the most relevant publications listed here: https://www.ltg.ed.ac.uk/software/geoparser/
- If you are interested in adapting the Geoparser for your own needs and would like to collaborate with us, do get in touch: (balex@inf.ed.ac.uk)

Questions or comments



Beatrice Alex, Edinburgh Language Technology Group (www.ltg.ed.ac.uk)

balex@inf.ed.ac.uk - @bea_alex - http://homepages.inf.ed.ac.uk/balex/

Questions from participants

- How long does the text have to be for the Geoparser to work well?

  - The vanilla download works most accurately with running English text. It works on individual sentences. Geo-resolution accuracy increases however if the Geoparser has access to more context. The Geoparser is not well suited to process large documents made up of several sub-texts, e.g. a journal made up of articles unless the articles are all related and contain similar locations. In the latter case it would be better to split the document into the articles first.

Questions from participants

- Is it possible to exclude particular names from the output ahead of geo-resolution?
  - This is not possible to do when treating the Geoparser as a blackbox. Currently we don't specify a command line option with a list of names to exclude. But this is something we will consider adding in future releases. If you become familiar with the tool you can edit the name lexicons located in `./lib/ner` and remove names that should not be tagged in the document.

Questions from participants

- Is it possible to specify/vary the context size used when disambiguating locations?
  - This is not possible in the default download. It considers all locations within the document when resolving a location in question.

Questions from participants

- Is it possible to geo-parse historical text using the Geoparser?

  - Yes. However, when using the Geoparser in combination with the GeoNames gazetteer some historical place names will not be identified as they are missing from the gazetteer. Also the Geoparser team can provide additional pre-processing to improve the quality of optical-character recognised output (e.g. to fix soft-hyphen splitting or to deal with the long “s” character). Those scripts are not distributed with the standard Geoparser distribution but you can contact us to get access to them.

Questions from participants

- Can the Geoparser output be visualised with another map interface (e.g. OpenStreetMap)

  - Yes. Once you have extracted the geo-location information from the *out.xml file(s) you can use it as input into any of your favourite mapping tool.