

# Notebook

September 15, 2014

```
In [1]: %pylab inline
        %load_ext autoreload
        %autoreload 2
```

Populating the interactive namespace from numpy and matplotlib

```
In [3]: import btb.utils.tools as btbttools
        import btb.utils.wikiquery as wq

        import mwclient
        from __future__ import division
```

## 1 Wikipedia page comparison

It is assumed that the length of a Wikipedia entry provides an indication of the level of interest users from that language have on the topic of that page.

Based on this assumption, comparing the length of an entry in English wikipedia and Dutch wikipedia will provide an indication of the level of interest english speakers and Dutch speakers have on this topic.

To verify this assumption, we will need access to both English and Dutch Wikipedias.

```
In [4]: enWiki = mwclient.Site('en.wikipedia.org')
        nlWiki = mwclient.Site('nl.wikipedia.org')
```

Given a page title (in English)...

```
In [5]: pageTitle = 'Ice skating'

        enPageTitle = pageTitle
        enPage = enWiki.Pages[enPageTitle]
        if enPage.redirect:
            enPage = enPage.redirects_to()
```

... we can find whether there is an entry for that title on Dutch wikipedia

```
In [6]: nlPageTitle = None
        langs = enPage.langlinks()
        for lang, langTitle in langs:
            if lang=='nl':
                nlPageTitle = langTitle
                print enPageTitle, ' ==> ', nlPageTitle
        if nlPageTitle!=None:
            nlPage = nlWiki.Pages[nlPageTitle]
```

```
Ice skating ==> Schaatsen
```

We can then calculate the ratio of length between Dutch and English entries

```
In [7]: nlPage.length/enPage.length
```

```
Out[7]: 0.8631531122217174
```

Other topics which are less typically Dutch, will have shorter entries and thus lower ratio.

```
In [8]: def getNLTitle(enPageTitle):
        enPage = enWiki.Pages[enPageTitle]
        if enPage.redirect:
            enPage = enPage.redirects_to()

        nlPageTitle = None
        langs = enPage.langlinks()
        for lang,langTitle in langs:
            if lang=='nl':
                nlPageTitle = langTitle
                print enPageTitle,' ==> ',nlPageTitle

        return nlPageTitle
```

```
In [9]: enPageTitle = 'Ice hockey'
        nlPageTitle = getNLTitle(enPageTitle)
```

```
enPage = enWiki.Pages[enPageTitle]
nlPage = nlWiki.Pages[nlPageTitle]
```

```
nlPage.length/enPage.length
```

```
Ice hockey ==> IJshockey
```

```
Out[9]: 0.21482600373590388
```

```
In [10]: nlPage.length,enPage.length
```

```
Out[10]: (18631, 86726)
```

## 2 Country considerations

However, because we are interested in translation from one country to another, comparing the page length may be misleading: entries in Dutch Wikipedia will be edited (mostly) by users from the Netherlands; however English Wikipedia will contain a more diverse mixture from various countries.

For this reason, perhaps it makes more sense to look at the countries which are making edits in both the English and Dutch Wikipedia. For English, we will concentrate on Canada; for Dutch, we will concentrate on the Netherlands.

```
In [11]: nlBots = wq.getAllBots(nlWiki)
        enBots = wq.getAllBots(enWiki)
```

```
{'augroup': 'bot'}
{'augroup': 'bot'}
```

```
In [12]: enPageTitle = 'Ice hockey'
        nlPageTitle = getNLTitle(enPageTitle)
```

Ice hockey ==> IJshockey

```
In [13]: ips, usrs, nrevs = wq.getContributionsForPage(enWiki, enPageTitle)
        byCC, conf, nIP, nUsrc, nBot, nUnkn = btbttools.prepareData(ips, usrs, enBots, lang='en')
        print enPageTitle,':',byCC['CA'],'Confidence: {:.4.2f}'.format(conf)
```

Ice hockey : 1762 Confidence: 0.60

```
In [14]: # Post NL user resolver
        ips, usrs, nrevs = wq.getContributionsForPage(nlWiki, nlPageTitle)
        byCC, conf, nIP, nUsrc, nBot, nUnkn = btbttools.prepareData(ips, usrs, nlBots, lang='nl')
        print nlPageTitle,':',byCC['NL'],'Confidence: {:.4.2f}'.format(conf)
```

IJshockey : 358 Confidence: 0.67

This one makes sense, not so many edits from NL on IJshockey

```
In [15]: enPageTitle = 'Ice skating'
        nlPageTitle = getNLTitle(enPageTitle)

        ips, usrs, nrevs = wq.getContributionsForPage(enWiki, enPageTitle)
        byCC, conf, nIP, nUsrc, nBot, nUnkn = btbttools.prepareData(ips, usrs, enBots, lang='en')
        print enPageTitle,':',byCC['CA'],'Confidence: {:.4.2f}'.format(conf)

        ips, usrs, nrevs = wq.getContributionsForPage(nlWiki, nlPageTitle)
        byCC, conf, nIP, nUsrc, nBot, nUnkn = btbttools.prepareData(ips, usrs, nlBots, lang='nl')
        print nlPageTitle,':',byCC['NL'],'Confidence: {:.4.2f}'.format(conf)
```

Ice skating ==> Schaatsen

Ice skating : 75 Confidence: 0.65

Schaatsen : 249 Confidence: 0.55

So maybe we should normalize these number of edits ? instead of the length ?

Use ratio of lengths as  $\text{length\_NL} * (\text{edits\_from\_NL} / \text{edits\_in\_NL\_WIKI}) / \text{length\_EN} * (\text{edits\_from\_CA} / \text{edits\_in\_EN\_WIKI})$  ?