

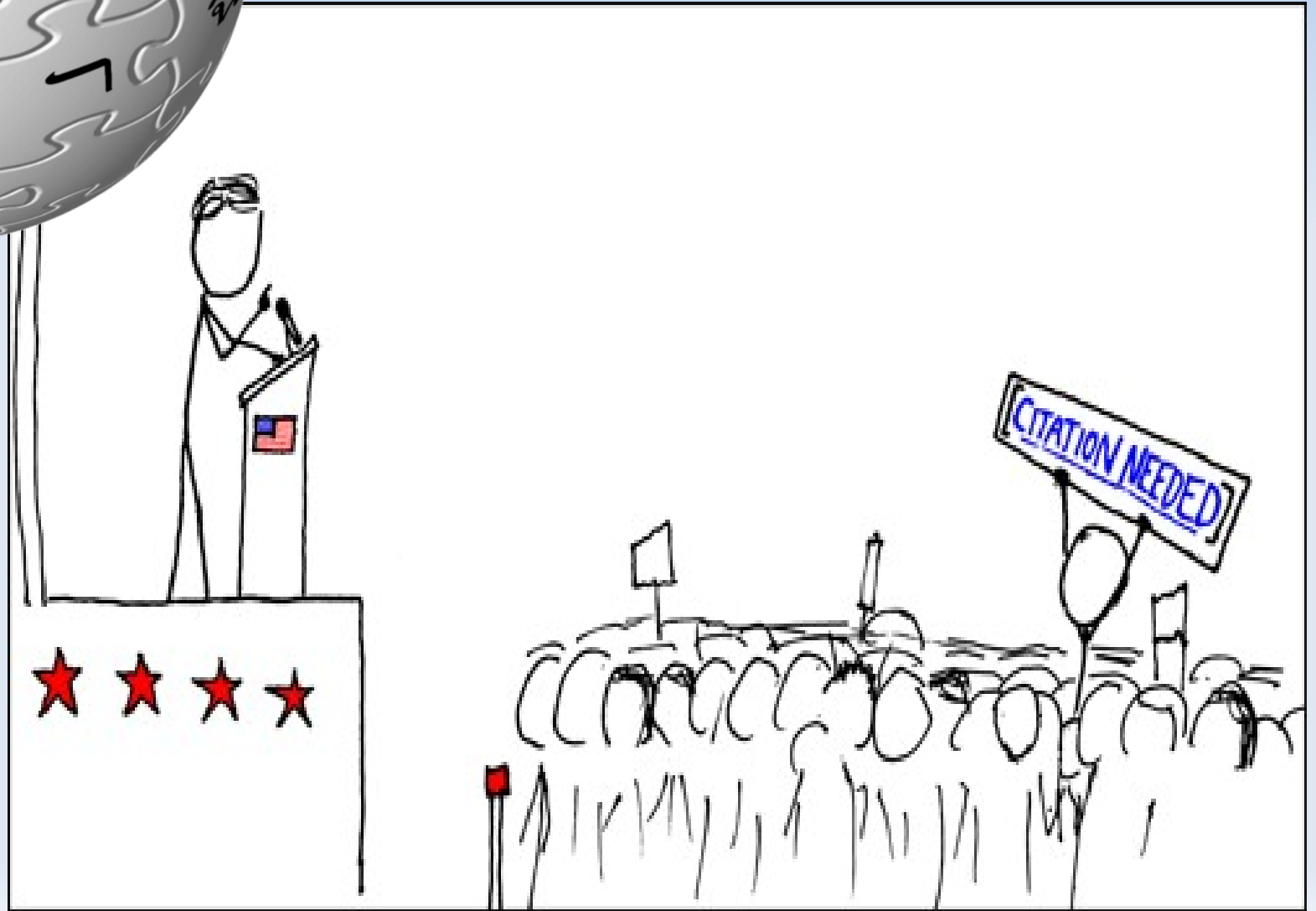
# **Interesting ways of using Wikipedia data**

# The presenters

- Prashanth Ellina
- Venkata Subramanian

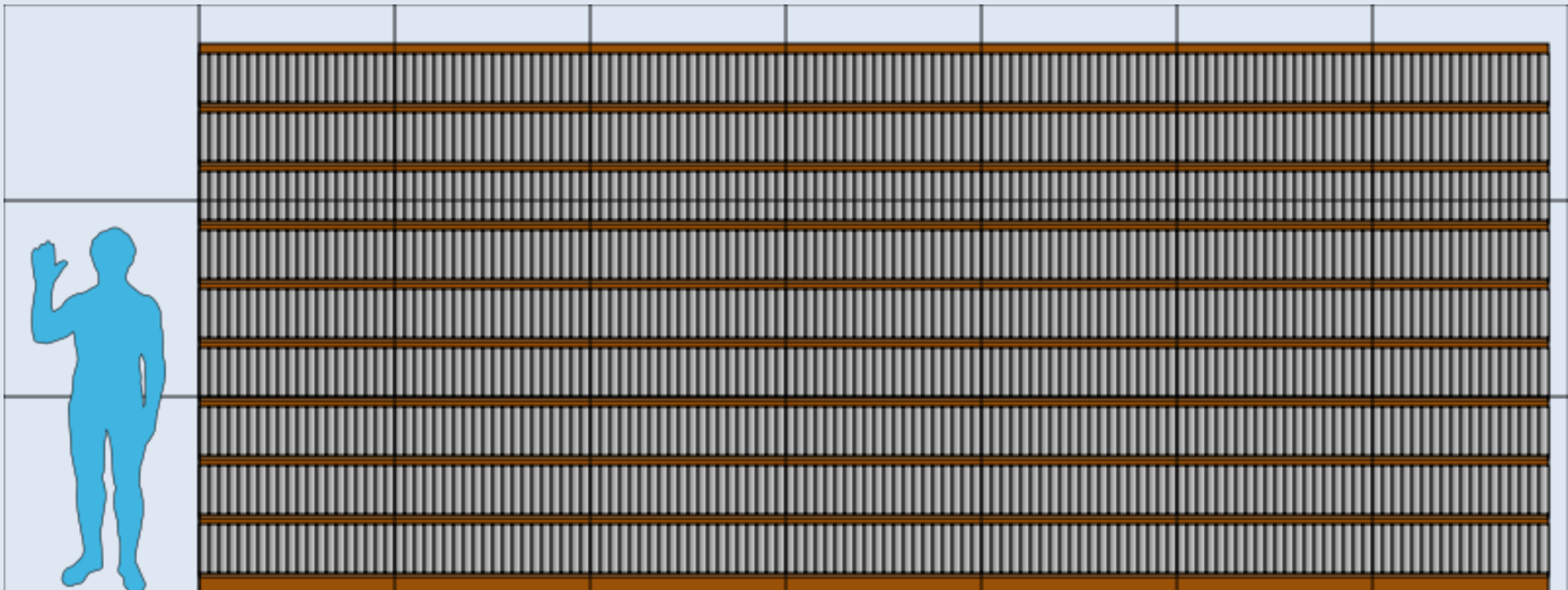


# Wikipedia?



# It is huge

- 2,000,000+ articles
- 136, 000, 000+ interpage links



# Corpus for data analysis

- Entities – Page titles
- Wiki graph – Page links
- Redirects and disambiguations
- Categorization – category links and lists
- Wiki text

# The Wikipedia graph



- Related articles in Wikipedia form clusters
- Sachin and Cricket

# Where is the Wikipedia data?

- Dumps available for download
- xml and sql files (5GB +)

# What is there in the dumps?

- page
- pagelinks
- text
- revision
- redirects
- categorylinks

# A Wikipedia article

- It has a unique page id
- also identified by page namespace + page title

# Page namespaces

- 0 – Articles
- 1 – Talk
- 2 – User
- 6 – Image
- 14 - Category

# Page table

- id
- namespace
- title
- is\_redirect

# Pagelinks

- from id
- to (namespace + title)

# Redirects

- from id
- to (namespace + title)

# text

- text id (this is not page id)
- the text

# revision

- revision id
- page id
- text id

# Setting up the database

- xml -> sql using xml2sql
- loading .sql files into mysql is almost straightforward

# Machine?

- Need a powerful machine
- recommended 4GB
- Memory bandwidth is vital
- Multiple HDD's increase speed.

**What did we use the data for?**

# News article clustering

- Current problem is "**Topic extraction**"

# Topic extraction

- Finding a "representative" set of Wikipedia articles for input text.

## **TENNIS: Federer falls to Djokovic in semis clash**

Roger Federer's invincibility was pierced, pummelled and finally pulled apart at Melbourne Park yesterday, the world number one brutally ejected from the Australian Open semi-finals by Novak Djokovic.

# Extracted topics

Roger Federer  
Australian Open  
Australia  
Melbourne  
Tennis  
Melbourne Park  
Novak Đoković  
Australian dollar

# Normalization and tokenization

tennis federer falls to djokovic in semis clash roger federer s  
invincibility was pierced pummelled and finally pulled apart  
at melbourne park yesterday the world number one brutally  
ejected from the australian open semi finals by novak djokovic

*apart, australian, brutally, by, clash, djokovic, ejected, falls, federer,  
finally, finals, from, invincibility, melbourne, novak, number, one,  
open, park, pierced, pulled, pummelled, roger, s,  
semi, semis, tennis, was, world, yesterday*

# Found titles

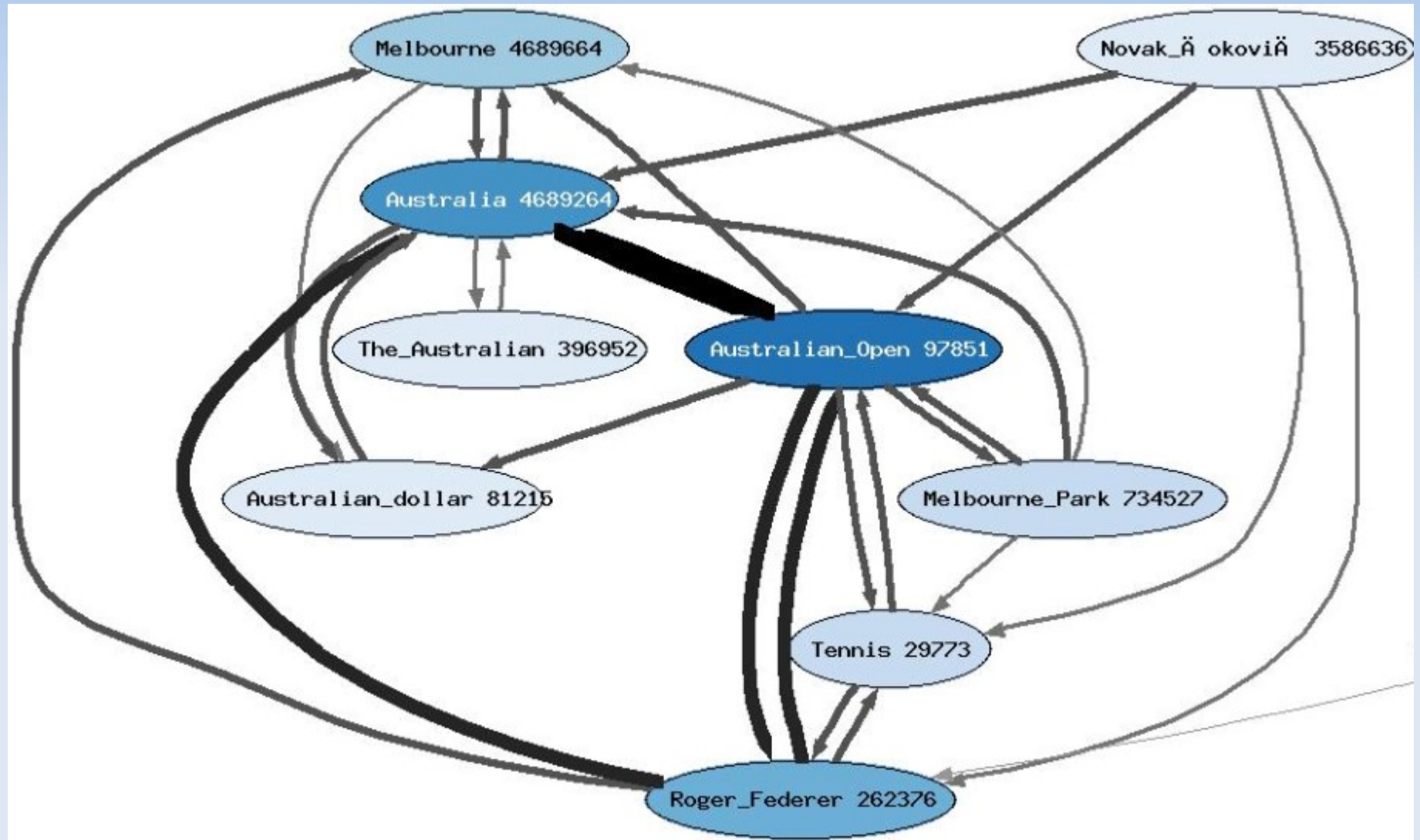
One'	one	8985819
(ONE)	one	6693468
-one	one	3675827
Australian	australian	92590
Australian_\$	australian	###
Australian_Open	australian open	97851
Clash	clash	78942
Clash!	clash	###
Djokovic	djokovic	###
Falls	falls	319513
Federer	federer	###
Finally	finally	###
Finally...	finally	###
Finals	finals	2793894
Invincibility	invincibility	2182190
Melbourne	melbourne	4689664
Melbourne_Park	melbourne park	734527
Novak	novak	1274827
Novak_Djokovic	novak djokovic	4316959
Number	number	21690

Number_One	number one	2450965
Number_one	number one	3848771
ONE	one	481849
One-	one	9360048
Pierced	pierced	6012155
Roger	roger	987632
Roger_Federer	roger federer	262376
Roger_federer	roger federer	8879742
SEMI	semi	3857595
Semi	semi	416399
Semi-	semi	1205636
Semi-finals	semi finals	2908791
Semis	semis	576417
SemiÄ	semi	4170795
Tennis	tennis	29773
The_Australian	the australian	396952
WAS	was	1715111
Was	was	1400413
Yesterday	yesterday	853973
,€one,€™	one	6471739

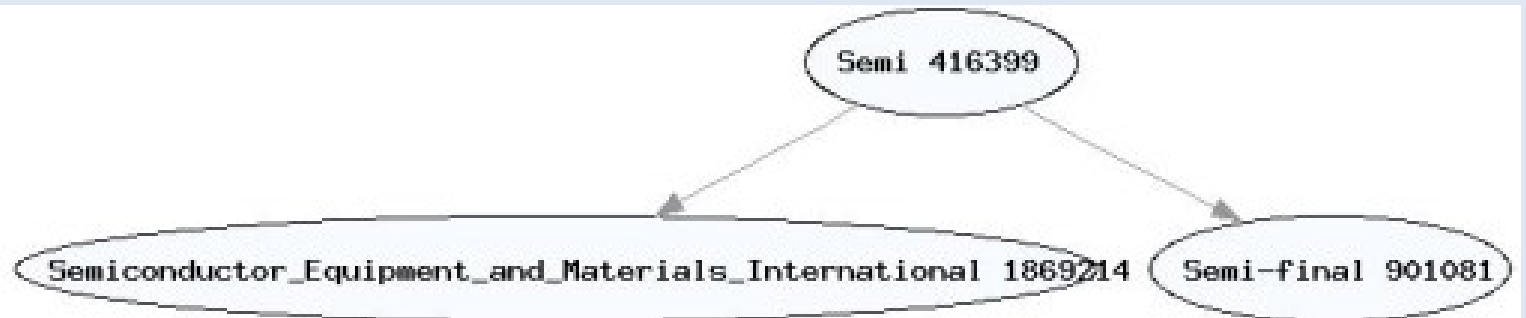
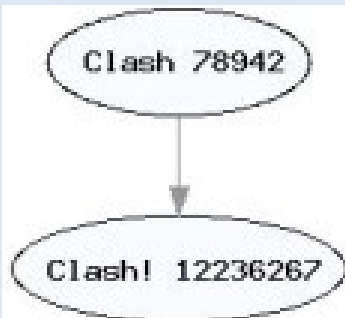
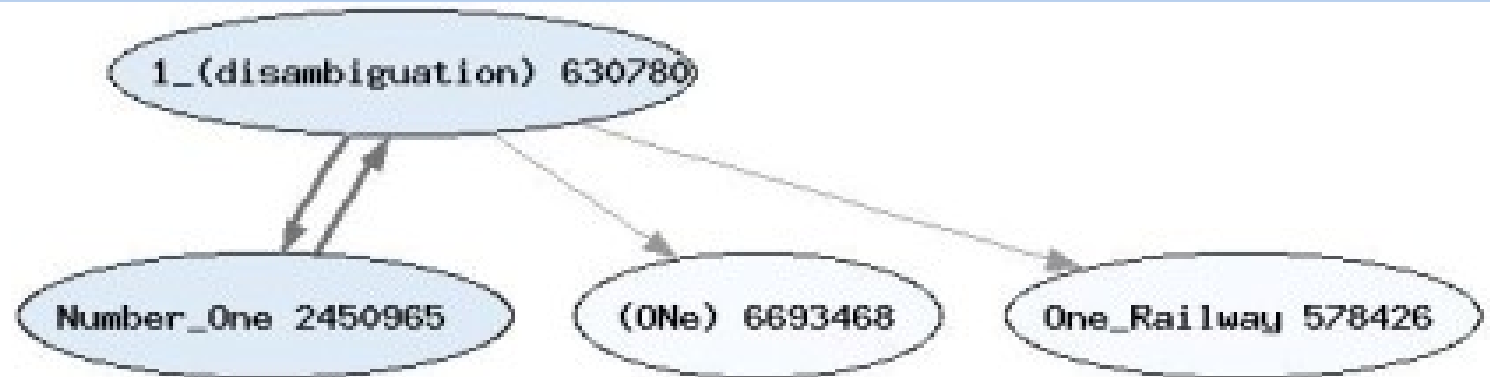
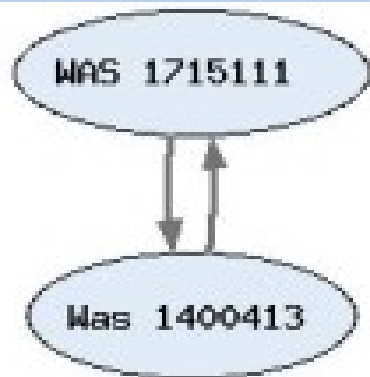
# Redirects

One' (8985819)	One_Railway (578426)
Australian (92590)	Australia (4689264)
Australian_\$ (12804226)	Australian_dollar (81215)
Djokovic (13038585)	ÄokoviÄ (10758924)
Falls (319513)	Falling (239099)
Federer (12620724)	Roger_Federer (262376)
Invincibility (2182190)	Invincible (106201)
Novak_Djokovic (4316959)	Novak_ÄokoviÄ (3586636)
Number_one (3848771)	Number_One (2450965)
ONE (481849)	1_(disambiguation) (630780)
One- (9360048)	1_(disambiguation) (630780)
Pierced (6012155)	Body_piercing (4867)
Roger_federer (8879742)	Roger_Federer (262376)
SEMI (3857595)	Semiconductor_Equipment_and_Materials_International (186921)
Semi- (1205636)	Numerical_prefix (1705121)
Semi-finals (2908791)	Semi-final (901081)
one™ (6471739)	One_Railway (578426)

# Big graph



# Disconnected nodes



# Node ranking

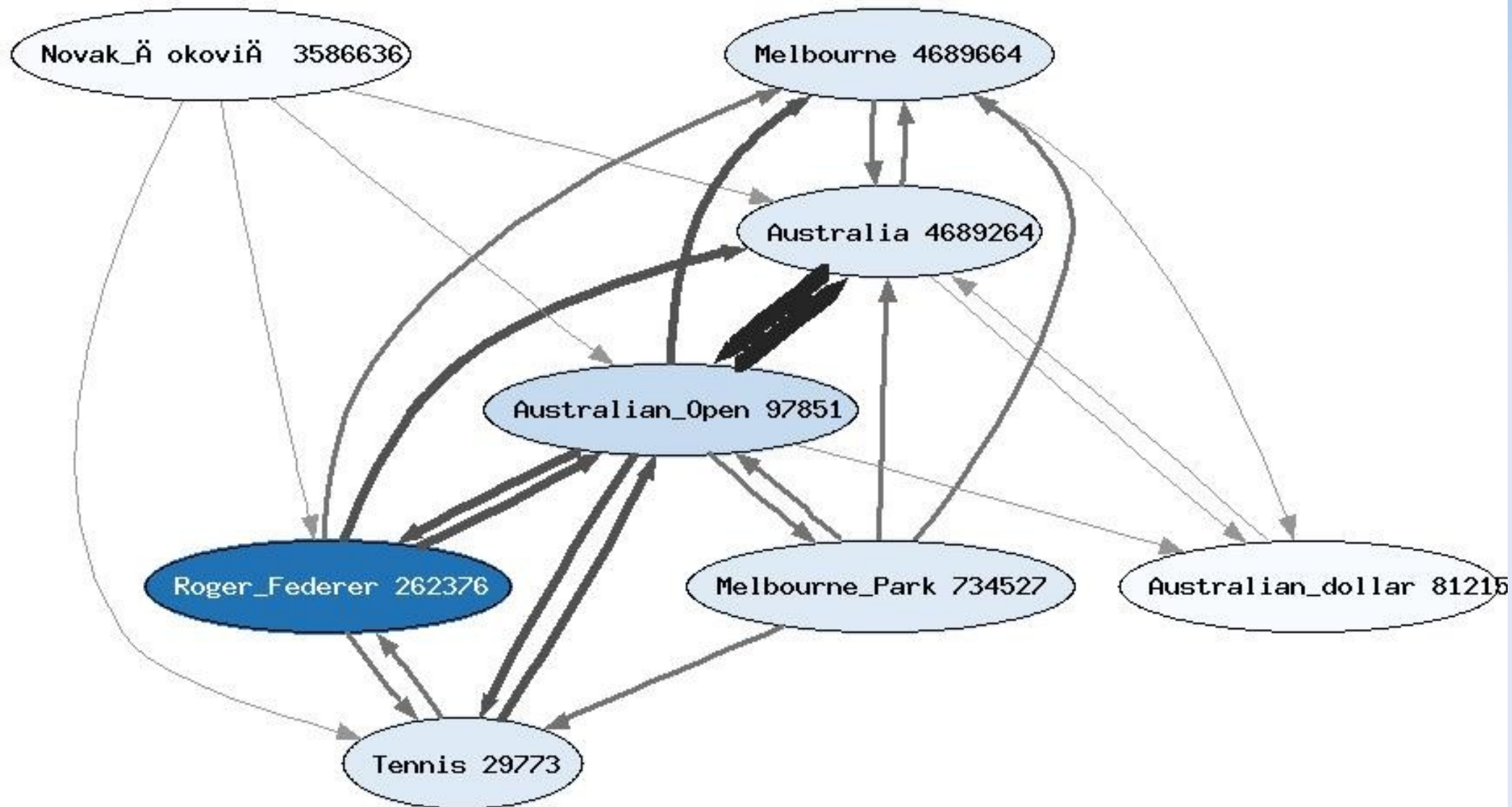
- $N_{\text{connections}} = \text{inlinks} + \text{outlinks}$
- $N_{\text{norm\_connections}}$
- $L_{AB} = N_{\text{norm\_connections}}^A * N_{\text{norm\_connections}}^B$
- $N_{\text{weight}} = \sum L_N$
- $N_{\text{norm\_weight}}$

# Choosing nodes

$$N_{\text{norm\_weight}} > 0.1$$

Australian_Open	97851	1 chosen
Australia	4689264	0.83 chosen
Roger_Federer	262376	0.53 chosen
Melbourne	4689664	0.35 chosen
Tennis	29773	0.33 chosen
Melbourne_Park	734527	0.25 chosen
Australian_dollar	81215	0.17 chosen
Novak_ÄokoviÄ	3586636	0.15 chosen
The_Australian	396952	0.03 rejected
Number_One	2450965	0.01 rejected
1_(disambiguation)	630780	0.01 rejected
WAS	1715111	0 rejected
Was	1400413	0 rejected
Clash!	12236267	0 rejected
(ONe)	6693468	0 rejected
Semiconductor_Equipment_and_Mater	1869214	0 rejected
Roger	987632	0 rejected
Semi-final	901081	0 rejected
One_Railway	578426	0 rejected
Semi	416399	0 rejected

# Sub graph



# Sub graph nodes after ranking

## Before match count adjustment

Australian_Open	97851	1
Australia	4689264	0.65
Roger_Federer	262376	0.45
Melbourne	4689664	0.25
Tennis	29773	0.25
Melbourne_Park	734527	0.14
Novak_ÄÄokoviÄ†	3586636	0
Australian_dollar	81215	

## After match count adjustment

Roger_Federer	262376	1
Australian_Open	97851	0.21
Australia	4689264	0.14
Melbourne	4689664	0.05
Tennis	29773	0.05
Melbourne_Park	734527	0.03
Novak_ÄÄokoviÄ†	3586636	0

# Issues

- Disambiguations are not handled
- Only one degree links are considered
- 0.1 threshold causes problems
- Too slow for real time operation

# Handling space/time issues

- about 180 seconds for average sized input
- most time spent in disk access by db

# Handling space/time issues

- Loading required wikipedia data into memory in python failed. Python taking about 32 bytes per integer.
- Memory requirement at the rate = 6GB+

# Handling space/time issues

- Moved data structures to C
- Interfacing with python using ctypes (highly recommended!)

# Handling space/time issues

- Memory requirement came down to 1GB
- 15 seconds for average sized input text
- We plan to try graph compression to reduce memory consumption further

# Other uses

- Understanding user interests
- Weird links extraction :)

# Weird Link extraction

# Links

- <http://blog.prashantheellina.com/tag/wikipedia/>
- <http://blog.prashantheellina.com/tag/python/>

prashantheellina@gmail.com  
venkatasubramanian@gmail.com

# Donate to Wikipedia!

- Very useful resource
- Start with a small sum

**We're done talking! Your turn :)**