IR Project 2013

Marco Cornolti
di.unipi.it/~cornolti

Relatedness measure: are two entities close?
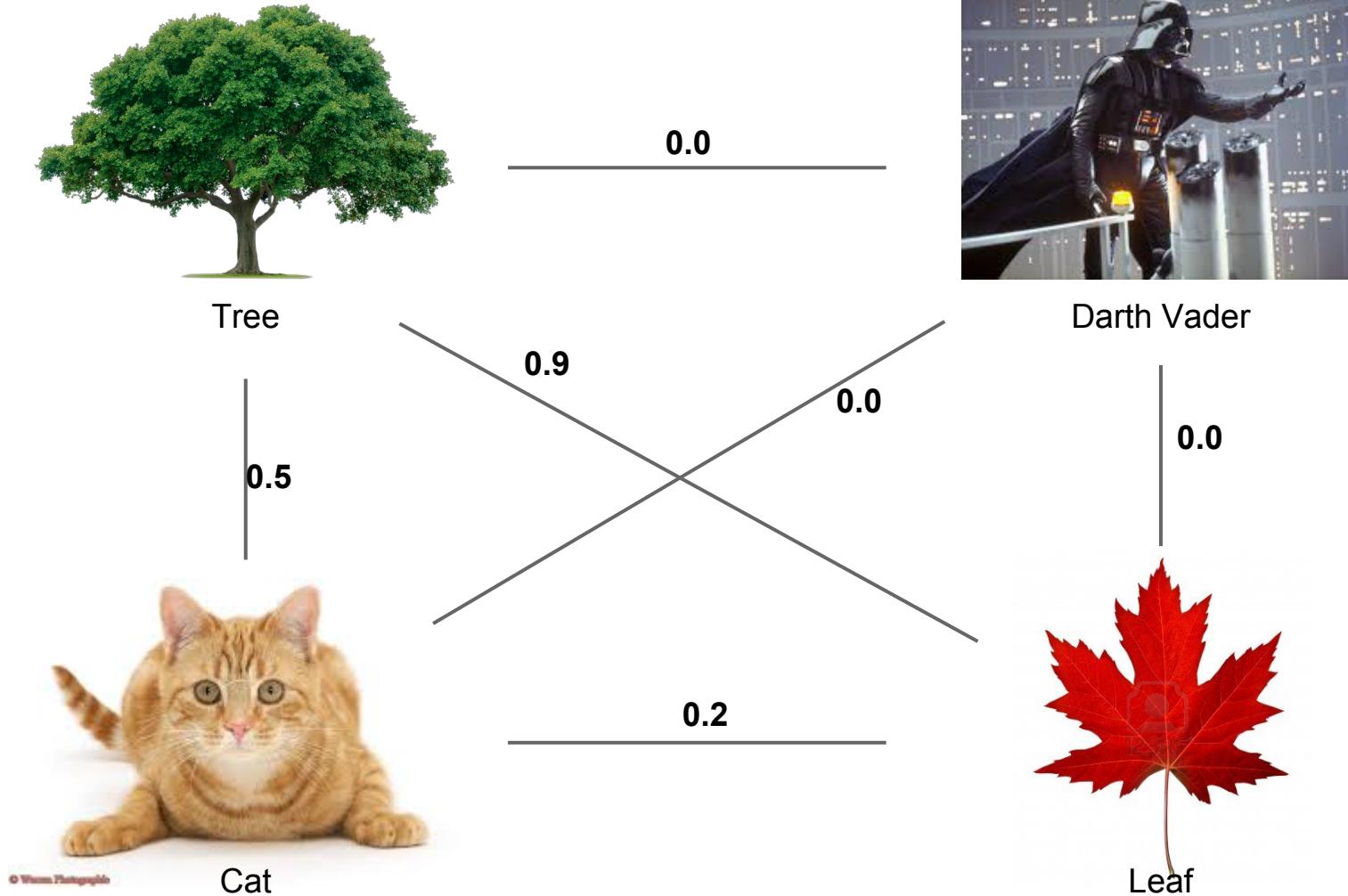
An example: in-link Jaccard
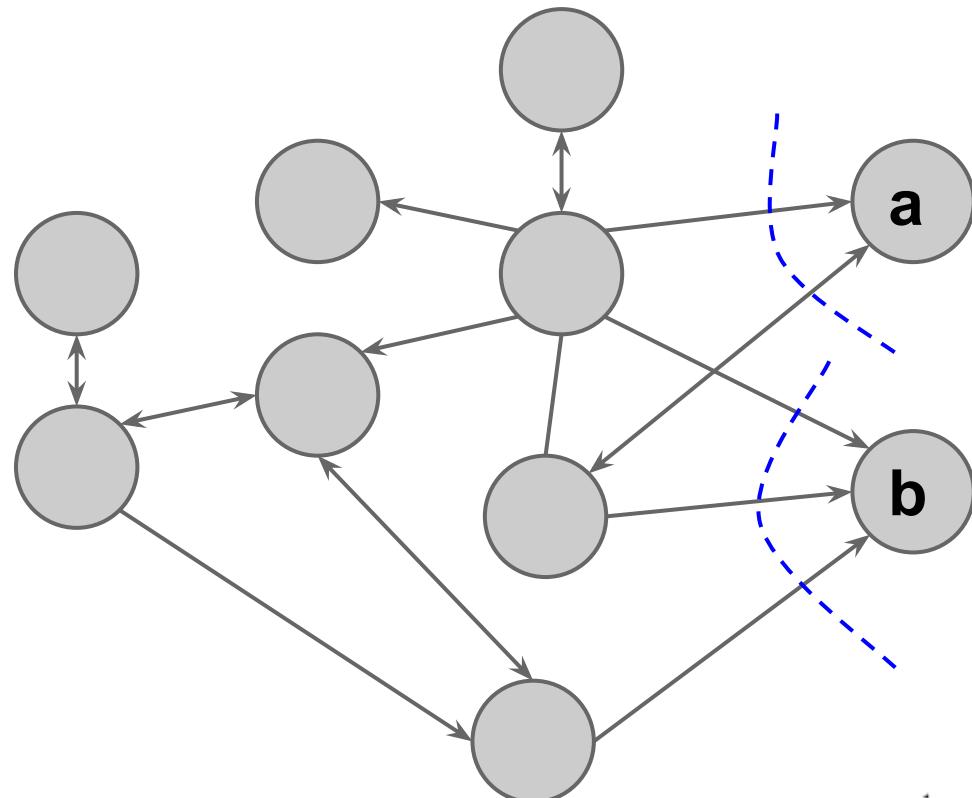

$$\frac{in(a) \cap in(b)}{in(a) \cup in(b)} = \frac{2}{2+3} = 40\%$$

Another example: Milne&Witten
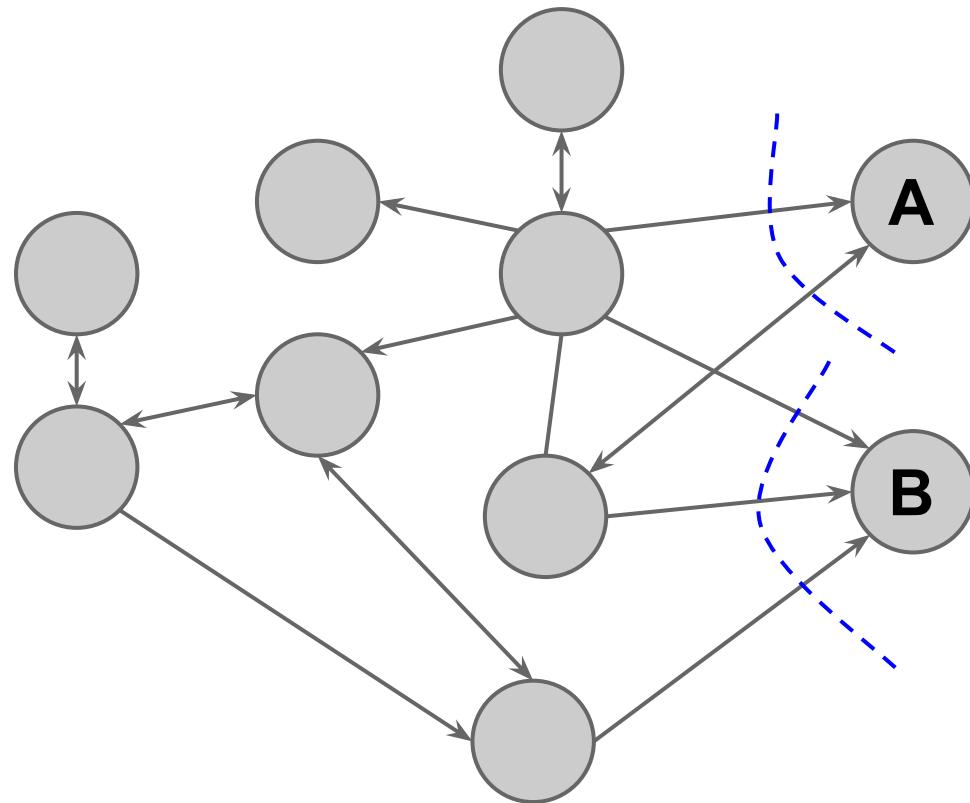
$$sr(a,b) = \frac{\log\big(\max\big(|A|,|B|\big)\big) - \log\big(|A \cap B|\big)}{\log\big(|W|\big) - \log\big(\min\big(|A|,|B|\big)\big)}$$

Things they do not consider

- direct mutual link

  A ⟷ B

- hubness of the pages
- size of the outlink star of the referring pages
- text similarity between pages

Our framework

test

BAT-Framework
test against
WikipediaSimilarity353

BAT-Framework
test against
AIDA/CO-NLL

applications

TagMe

launch the expriments

IRProjectExperiment
(runs experiments)

relatedness function

Your relatedness
function goes here

helper functions

IRProjectHelper
(helper functions)
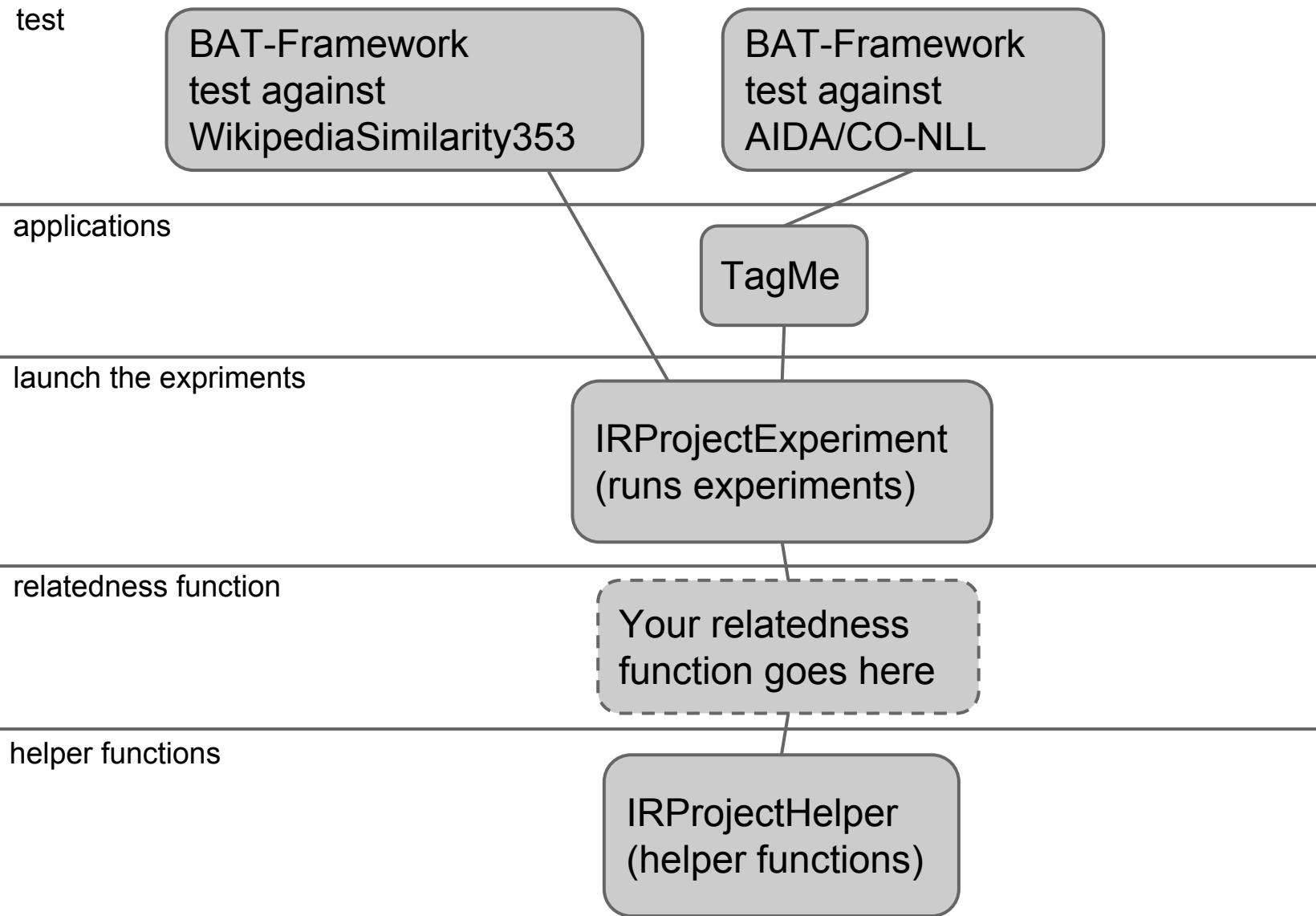
Helper functions (`IRProjectHelper`)

```
public static int[] getInlinks(int page_id);
public static int[] getOutlinks(int page_id);
public static int TitleToId(String title);
public static String getCategoryTitle(int catId);
public static IntSet getAllWids();
public static boolean isDisambiguation(int pageId);
public static boolean isNormalPage(int page_id)
public static boolean isPerson(int pageId);
public static int[] getCategories(int pageId);
public static int dereference(int pageId);
public static float linkProbability(string anchor);
public static float commonness(string anchor, int pageId);
```

...see the Javadoc

Example implementation: Jaccard

```java
public class Jaccard extends RelatednessMeasure {
    public Jaccard(String lang) {
      super(lang);
    }

    @Override
    public float rel(int entity1, int entity2) {
      int[] inA = IRProjectHelper.getInlinks(entity1);
      int[] inB = IRProjectHelper.getInlinks(entity1);
      HashSet<Integer> intersection = getIntersection(inA, inB);
      HashSet<Integer> union = getUnion(inA, inB);
      return (float) intersection.size() / (float) union.size();
    }
//...
}
```

How to run the experiments: `Main.java`

```java
public class Main {
  public static void main(String[] args) throws Exception {
    TagmeConfig.init("/home/irproject/config.xml");
    String groupName = "";
    String groupPw = "";
    RelatednessMeasure rel = null;
    IRProjectExperiments.
        launchRelatednessExperiment(groupName, groupPw, rel);
    IRProjectExperiments.
        launchTagMeExperiment(groupName, groupPw, rel);
  }
}
```

These will perform the experiments and publish the results.

On-line ranking

Running the experiment

ssh

scp

sshfs

java

javac

fiveg

Commands

```
ssh groupN@ferrax-2.itc.unipi.it

export IRLIB=../irproject/lib/*:../irproject/tagme_libs/*: /
../irproject/tagme_preproc_lib/*

javac -cp $IRLIB -d bin/ src/irproject/*

java -Xmx8g -cp $IRLIB:bin/ irproject.Main
```

Deadline & details

- To get un&pw, email me.
- Deadline is on **Feb 9, 12:00**. We will check your code and look at the rankings.
- Please leave in your home directory **the whole** code **only**. The code must include a `Main.java` file that launches the experiments.
- You will make a pitch (5min presentation) on **Feb 10, 9:30 @ Aula Seminari Ovest**.

13,72M years ago   8 Jan   29 Jan   9 Feb, 12 am   10 Feb, 9:30 am

big bang   1st test   2nd test   deadline   pitches
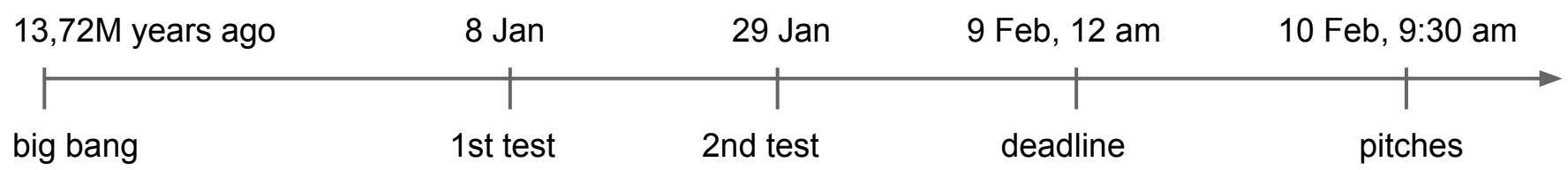
Suggestion

- Do some brainstorming before starting to implement
- You may ask for feedback at any time by e-mail or in my room (let me know in advance)
- Expand your horizon: you can implement more functions than these
- We encourage good ideas rather than good results
- Numbers are big: do not engineer, but be careful with complexity.