Appunti progetto di WSIE:

Bisogna utilizzare il dataset WikiMID: è un dataset TSV, ovvero Tab Separated Values. Il primo punto è dunque quello di leggere il dataset in modo coerente con quello che bisognerà fare dopo. Questo:

* contiene circa 500000 utenti (sia inglesi che italiani)
* associa un insieme di interessi estratti dai loro messaggi o dalla loro lista degli amici (selezionando quegli amici che indicano più un certo tipo di interesse che un vero e proprio rapporto di amicizia)
* associa una pagina di Wikipedia ad ogni interesse

ATTENZIONE: BISOGNA USARE SOLO IL DATASET INGLESE!

# Task1: trova categorie

Per ogni utente u\_i, dati i suoi interessi mappati in pagine di Wikipedia, trova una serie di CATEGORIE che rappresentino una sintesi degli interessi che ha mostrato. In generale le categorie dovrebbero essere meno degli interessi.

Nell'esempio fornito notare che NON c'è una corrispondenza di coppia tra le pagine di wikipedia e gli interessi. Questi sono estratti collettivamente dall'intero insieme delle wikipagine per ciascun utente.

Per riuscire a fare questa cosa si possono utilizzare molte risorse semantiche: le categorie di Wikipedia, DBPedia, Babelnet e qualsiasi altro metodo inventato o trovato in letteratura e non ci si aspetta che sia innovativo.

# Task2: genera cluster (serve task1 –forse--)

Bisogna generare cluster degli utenti simili (ad esempio che hanno interessi simili) utilizzando qualsiasi metodo di community detection visto a lezione.

# Task3: valuta cluster (serve task2)

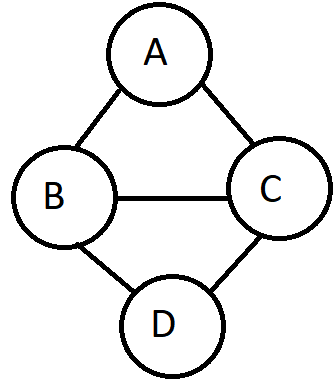
Utilizzare un qualsiasi metodo semplice per valutare i cluster ottenuti al punto precedente (per esempio la distanza media tra i membri di un cluster e tra i membri di cluster diversi: ovviamente i membri dello stesso cluster devono essere più simili gli uni agli altri, rispetto a qualsiasi due elementi che appartengono a cluster diversi)

# Task4: usa API su S21(ind.) e metti user in cluster (forse serve 2.)

In un altro file (S21.tsv) ti vengono dati 1500 ID di Twitter. Utilizzando le API di Twitter, scarica le informazioni su questi 1500 profili e sulle loro amicizie, quindi tenta di associare ogni user u\_j con il cluster di altri user più simili a lui. Spiega il metodo di similarità adottato.

# Task5: usa API su S22 (serve task4), consiglia item (--forse-- indipendente)

In un altro file ancora (S22-preferences.tsv) vengono dati altri 500 users, per cui si ha l'ID e la lista degli items preferiti in termini di pagine di Wikipedia. Inoltre, per ogni user viene fornita una lista di 6 pagine Wikipedia (dalla lista del file S23.tsv). Devi raccomandare ad ogni user 3 dei 6 item proposti, scartando gli altri 3. Per farlo in teoria dovresti definire ed implementare alcuni algoritmi che facciano il rank dei 6 items in accordo agli interessi dell'utente precedentemente dedotti. Spiegare il metodo utilizzato per decidere e quali raccomandazioni sembrano essere più idonee per gli interessi dell'utente.

Descrizione dataset

# WikiMID

Mappa : **Id\_user Id\_tweet Id\_interest Interest\_url**

* A-friend-based\_dataset.tsv: Id\_user Id\_friend
* B-friend-based\_interest.tsv: Id\_friend Wiki-page
* C-message-based\_dataset.tsv: Id\_user Id\_tweet Id\_interest interest\_url
* D-message-based\_interest.tsv: Id\_interest platform\_type wiki\_page

Come organizzarlo in Java?

Proposta: fare una classe **user** che contenga come campi: l'id, la lista degli amici (references ad altri user o ad "item"?), la lista dei tweet. In questo modo gli "item" e i tweet sono altre due classi : **item** ha come campo la wiki-page corrispondente e la lista degli interest inerenti. Invece **tweet** ha come campo l'url e la lista degli interest corrispondenti.

Si potrebbe pensare di fare una classe apposta per **interest,** ma ora come ora penso ci andrebbe solo la il nome dell'interesse. E' comunque utile, pur essendo na classe mezza vuota, per implementare metodi propri degli interessi e magari si scopre che è comodo aggiungervi roba.