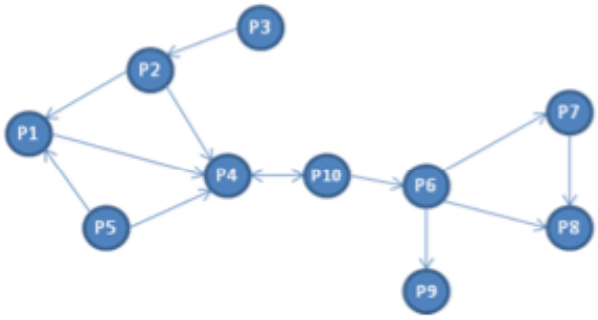
**NOME**: **COGNOME**:

Esercizio 1 – 12 punti
(**Page Rank**)




Sia data la rete con i collegamenti in figura, riassunti qui sotto:
P1 -> P4
P2 -> P1, P2 -> P4
P3 -> P2,
P4 <-> P10
P5 -> P4, P5 -> P1,
P6 -> P7, P6 -> P8, P6 -> P9,
P7 -> P8
P10 -> P6
Partendo da pesi iniziali tutti pari ad 1, calcolate le prime 2 iterazioni di Page Rank usando la formula:

$$PR(u) = d \sum_{all\ v\ links\ to\ u} \frac{PR(v)}{N_v} + (1-d)$$

con d=0,85

Esercizio 2 (6 punti)
**Inverted Index**
Dati i seguenti documenti:
D1: You say goodbye, I say hello
D2: You say stop, I say go
D3: Hello, hello, you say goodbye
D4: I say yes, you say no
Costruite un inverted index

Esercizio 3 (12 punti)
**SVD**
Sia L una matrice sulle cui righe compaiono i vettori dei documenti (le cui celle rappresentano il peso del termine j nel documento i). La matrice L tende ad essere molto sparsa (molti elementi pari o vicini a zero). Il metodo SVD è un metodo algebrico sul quale si basa la tecnica di IR detta *latent semantic indexing (LSI)*.

- Illustrare il metodo SVD e l'algoritmo LSI in maniera formale,
- Spiegare come mai la riduzione di rango operata da SVD equivale a raggruppare (effettuare un clustering de)i termini attorno a delle direzioni principali,
- Chiarire cosa si intenda per "direzioni principali" in relazione al concetto di *eigenvectors*,
- Cosa sono da un punto di vista algebrico e cosa rappresentano dal punto di vista dell'IR le matrici V', S' e U ' ottenute a seguito della decomposizione e riduzione di rango di L?

## Soluzioni:

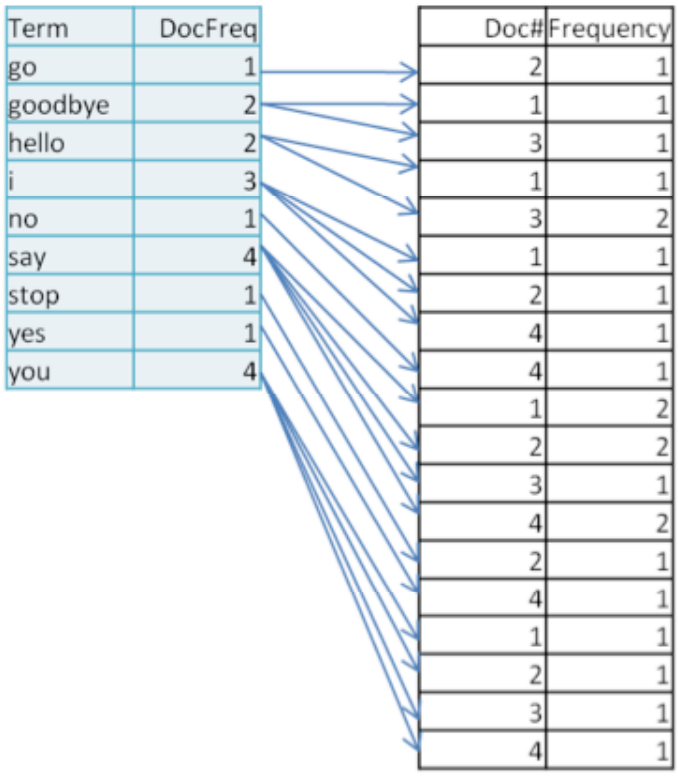
Esercizio 1: Esempio di calcolo di un passo per un nodo (P8):

PR(P8) = 0.85*(PR(P6)/3+PR(P7)) +0.15

Le prime 3 iterazioni:

| Iteration | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 |
|---|---|---|---|---|---|---|---|---|---|---|
| initial | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1.0000 | 1.0000 | 0.1500 | 2.2750 | 0.1500 | 0.5750 | 0.4333 | 1.2833 | 0.4333 | 1.0000 |
| 2 | 0.6388 | 0.2775 | 0.1500 | 1.9138 | 0.1500 | 0.5750 | 0.3129 | 0.6813 | 0.3129 | 2.0838 |
| 3 | 0.3317 | 0.2775 | 0.1500 | 1.7602 | 0.1500 | 1.0356 | 0.3129 | 0.5789 | 0.3129 | 1.7767 |

Esercizio 2: L'inverted index è mostrato in figura, le frecce sono puntatori.



| Term | DocFreq |
|---|---|
| go | 1 |
| goodbye | 2 |
| hello | 2 |
| i | 3 |
| no | 1 |
| say | 4 |
| stop | 1 |
| yes | 1 |
| you | 4 |

| Doc# | Frequency |
|---|---|
| 2 | 1 |
| 1 | 1 |
| 3 | 1 |
| 1 | 1 |
| 3 | 2 |
| 1 | 1 |
| 2 | 1 |
| 4 | 1 |
| 4 | 1 |
| 1 | 2 |
| 2 | 2 |
| 3 | 1 |
| 4 | 2 |
| 2 | 1 |
| 4 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |

Dove **docfreq** è la document frequency, cioè la frequenza di un termine nella collezione usata per il calcolo di *idf*, e **Frequency** è la frequenza del termine nel documento.

Esercizio 3: vedere appunti su LSI e articoli sul sito del corso.