

# GraphCorpus

## Un Outil d'Aide à l'Analyse de Corpus

### Problématique

#### Analyse automatique de corpus (dépêches)

- Comment **regrouper les documents qui traitent d'un même événement** sans effectuer de traitements lourds ?

Ce type d'analyse présente un intérêt pour :

- La **veille d'informations** ( cartographie des dépêches fournies par un média sur une période ou un événement donné )
- La **synthèse automatique** ( production de résumés à partir de regroupements de documents traitant d'un même événement )

### Les Réalisations

- Un corpus **annoté en entités nommées** est analysé et représenté sous la forme d'un graphe pondéré :
  - chaque noeud est un document ;
  - le poids des liens est établi grâce à la mesure de similarité exposée ci-contre.
- A partir de cette analyse, **GraphCorpus** fournit une **cartographie** à l'aide d'une méthode SOM (Meyer 1998), grâce à GraphExplore (<http://graphexplore.cagp.duke.edu>)

### Similarité entre documents

- Nouvelle mesure fondée sur Jaccard, pondérée par le poids des entités. Le poids d'une entité est uniquement fonction de son type (date, lieu, personne, ...). Cela permet de rendre compte du caractère discriminant d'un type spécifique d'entité.

$$S(i, j) = \frac{N_{1,1}(i, j)}{N_{0,1}(i, j) + N_{0,1}(j, i) + N_{1,1}(i, j)}$$

$$N_{0,1}(i, j) = \sum_{k=0}^n (\text{poids}(e_k) | e_k \in j, e_k \neq i)$$

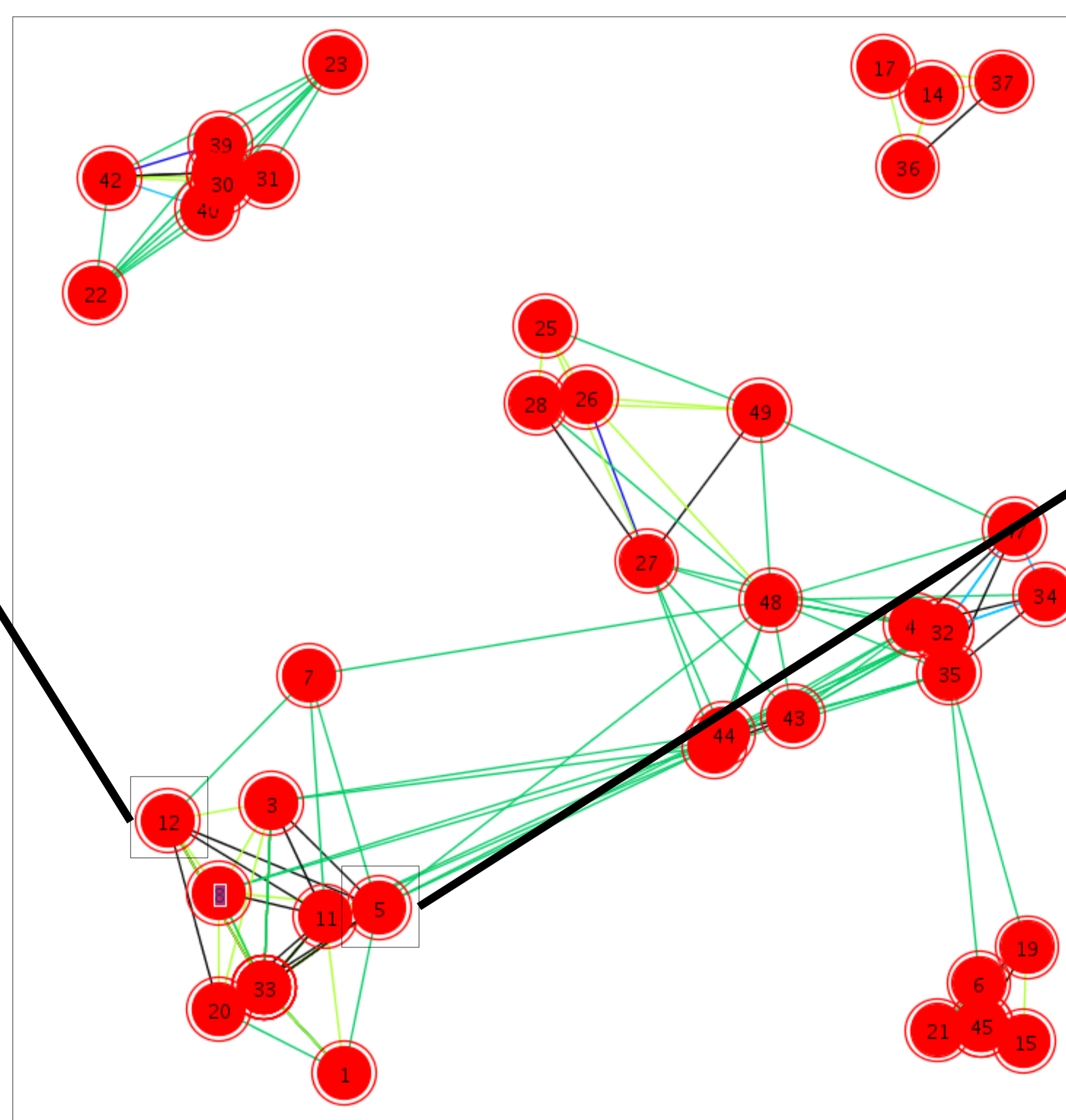
$$N_{1,1}(i, j) = \sum_{k=0}^n (\text{poids}(e_k) | e_k \in i, e_k \in j)$$

La marche de l'ancienne majorité est une «provocation d'émeutes», selon **Gaston Flossé** [16/10/2004 09:27] **PARIS** (AP) -- Le sénateur **Gaston Flossé** (UMP), candidat à la présidence de la **Polynésie** française a qualifié vendredi la marche qui était organisée ce samedi par l'ancienne majorité autour d'**Oscar Temaru** de «soulèvement populaire» et de «provocation d'émeutes». Invité sur le plateau de **RFO-Polynésie**, il a estimé qu'il s'agissait pour le président destitué **Oscar Temaru**, dont le gouvernement a été renversé samedi dernier par une motion de censure, «de faire une démonstration de son option pour l'indépendance». «C'est une marche anti-démocratique», a-t-il martelé. La «marche pacifique», organisée à **Paitia** par les partis politiques soutenant **Oscar Temaru**, et à laquelle doit participer une délégation du **PS** conduite par l'ancien secrétaire d'Etat à l'Outre-mer **Christian Paul**, pourrait réunir autour de 10.000 personnes, selon les médias locaux. Les manifestants devaient réclamer la dissolution de l'Assemblée de la **Polynésie** française, refusée à **Oscar Temaru** par deux fois par la ministre de l'Outre-mer **Brigitte Girardin**. Une pétition circulera dans les rangs des marcheurs pour appuyer également cette demande. Dans son entretien à **RFO-Polynésie**, **Gaston Flossé** a par ailleurs évoqué son retour au pouvoir, qu'il avait occupé quasiment sans interruption depuis **xxxx/1983** avant d'être remplacé par l'indépendantiste **Oscar Temaru** après les élections de **xx/05/2004**. «Je demande aux Polynésiens de me donner leur confiance», a-t-il déclaré, avant d'annoncer: «Au bout d'un an, je ferai mon bilan et s'il n'est pas suffisant, si nous n'avons pas réussi à relever le pays, je demanderai une motion de confiance à l'assemblée». **Gaston Flossé**, âgé de 73 ans, a été jusqu'ici toujours très évasif sur son retrait politique, et n'a jamais désigné nommément son successeur. Lundi, lors du comité central de son parti **Tahoeraa Huiraatira**, où il a été désigné à l'unanimité candidat à la présidence de la **Polynésie** française, il a toutefois assuré que le mandat qu'il briguerait serait «le dernier». AP l'p/nc

Les deux documents (ci-dessus et à droite) sont des dépêches issues d'un corpus de l'AFP.

Les parties sur-lignées correspondent aux entités nommées annotées dans le texte.

La proximité de contenu réelle entre ces deux documents est visible sur le graphe (les deux documents font partie d'un même groupe, un lien très fort – en noir – les unissant)



**Polynésie**: **Jack Lang** dénonce «des méthodes de république bananière» [17/10/2004 16:24] **PARIS** (AP) -- L'ancien ministre socialiste **Jack Lang** a demandé dimanche à **Jacques Chirac** de dissoudre l'Assemblée de **Polynésie** française pour mettre fin à «une sorte de coup d'Etat légal» qui a selon lui été «ordonné par les gens du pouvoir parisien» au profit de l'ancien président polynésien **Gaston Flossé** (UMP). «L'ensemble de l'exécutif français a voulu annihiler le résultat du vote populaire en **Polynésie** par lequel un nouveau gouvernement a été désigné (et) M. **Gaston Flossé** a été destitué», a accusé **Jack Lang** sur **Radio3**. «Ce sont des méthodes de république bananière», a-t-il soutenu. La veille, plus de 15.000 personnes ont manifesté à **Papeete** pour réclamer la dissolution de l'assemblée et la tenue de nouvelles élections. Le **09/10/2004**, le gouvernement de l'indépendantiste **Oscar Temaru** élu à la présidence grâce à la courte majorité acquise aux élections de **xx/05/2004** avait été renversé à la suite de l'adoption d'une motion de censure à l'assemblée. «Pour la paix civile, il faut dissoudre l'assemblée de **Polynésie** et demander au peuple d'être l'arbitre de cette situation», a demandé **Jack Lang** pour qui le «président de la République française» doit prendre cette décision. La ministre de l'Outre-mer **Brigitte Girardin** a refusé à plusieurs reprises de dissoudre l'assemblée de la **Polynésie**. Le sénateur **Gaston Flossé**, qui a dirigé quasiment sans interruption la **Polynésie** française depuis **xxxx/1982**, est à nouveau candidat à la présidence. AP l'p/nc

Ce schéma présente une cartographie de 55 documents. Chaque noeud correspond à un document.

Les documents qui présentent un contenu proche sont regroupés géographiquement dans la mesure du possible.

La couleur d'un lien dénote la similarité entre les deux documents qu'il lie. Plus la couleur est claire, plus la similarité est faible.

### Perspectives

- Etablir d'autres mesures de similarité entre documents
- Réaliser un **clustering automatique** du corpus avec différents systèmes d'apprentissage sur différents types de données (données brutes, matrice de distance...) afin d'évaluer leur pertinence dans un tel cadre.
- Générer un **résumé automatique** pour chacun des clusters repérés.