Etude d'un graphe d'argumentation

Perreard Antoine/Mekki Yasmina/Du Yihan/Zhang Yilun

EISTI

perreardan@eisti.eu

April 19, 2017

Contexte

- De plus en plus de données (réseaux sociaux,...)
- Utilisation de ces données pour comprendre et prévoir une situation (élection)
- Analyser les idées politiques et les comparer
- Deux possibiltés: Fond / Forme

Sommaire

Objectifs

Notre objectif est de pouvoir représenter des arguments sur un sujet et de pouvoir les étudier:

- Evaluer la qualité d'un argument
- Trouver les noeuds les plus controversés
- Regrouper des acteurs ayant des arguments similaires
- Prédiction de liens via machine learning

Extraction des données

Création des données:

- Analyse des débats et extraction des arguments "pour" ou "contre" un sujet donné (immigration, économie, ...)
- Retranscription directe sous le format gml
- Modélisation des noeuds (arguments) et des interactions entre eux par des liens (attaques/contre attaques).

Modélisation

Les noeuds centraux

Un sujet sous forme de question fermée: pour ou contre

Les noeuds

Un acteur
Son bord politique
Son contenu
Une position vis a vis du sujet

Les liens

liens d'attaque -
Liens d'approbation +
liens entre acteurs (attaque) ou vers le sujet directement (attaque ou approbation)

Exemple de graphe

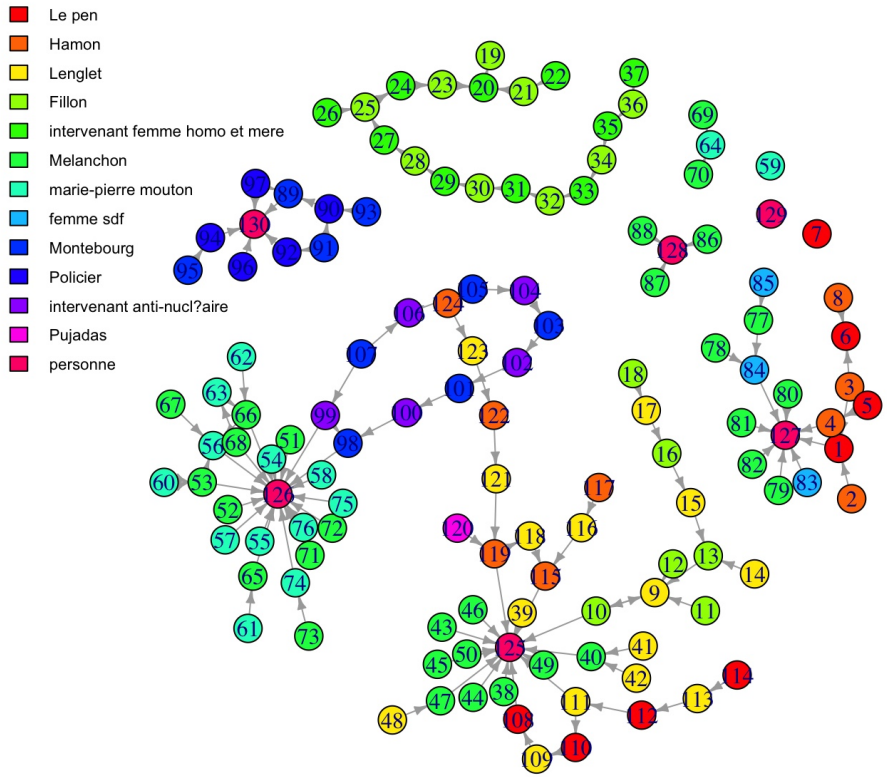


Figure: Répartition des noeuds selon les acteurs

Création du graphe sous neo4j

- Transformer les données: GML=>CSV=>Cypher de Neo4j
- Pourquoi NEO4J?
  - Débuter sur de bonne base (infrastructure)
  - Reformer le graph rendant les requêtes plus facile comme Shortest Path
  - Augmenter la capacité de stockage
- Possibilité de co-opérer avec apache beam dans l'avenir

Statistique descriptives

Pour le graphe complet:

- nombre d'idées = 6 (économie,demanteler les bases nucléaire,immigration,voile,famille, service national)
- nombre de noeuds = 130
- nombre de liens = 138
- densité = 0.016
- diamètre = 16

Statistique descriptives : centralité

Utilisation des fonctions :

- centralité de proximité

Conclusion: Voir les noeuds les plus centraux :

- 0.011059 : Intervention de l'état
- 0.011039 : Moins de bureaucratie
- 0.011034 : Besoin d'un revenu universel pour faire une transition et aider
- 0.011031 : Concurence avec d'autres pays. Risque de perdre des marchées et donc de la croissance
- 0.011028 : Partage du travail à cause de sa rarification dû à la numérisation
- 0.011026 : Fabrication d'eoliennes en France pour une relance économique
- 0.011025 : Relocalisation des entreprises pour créer de l'emploi en France
- 0.011024 : Baisser les impôts des sociétés

Statistique descriptives : centralité

Utilisation des fonctions :

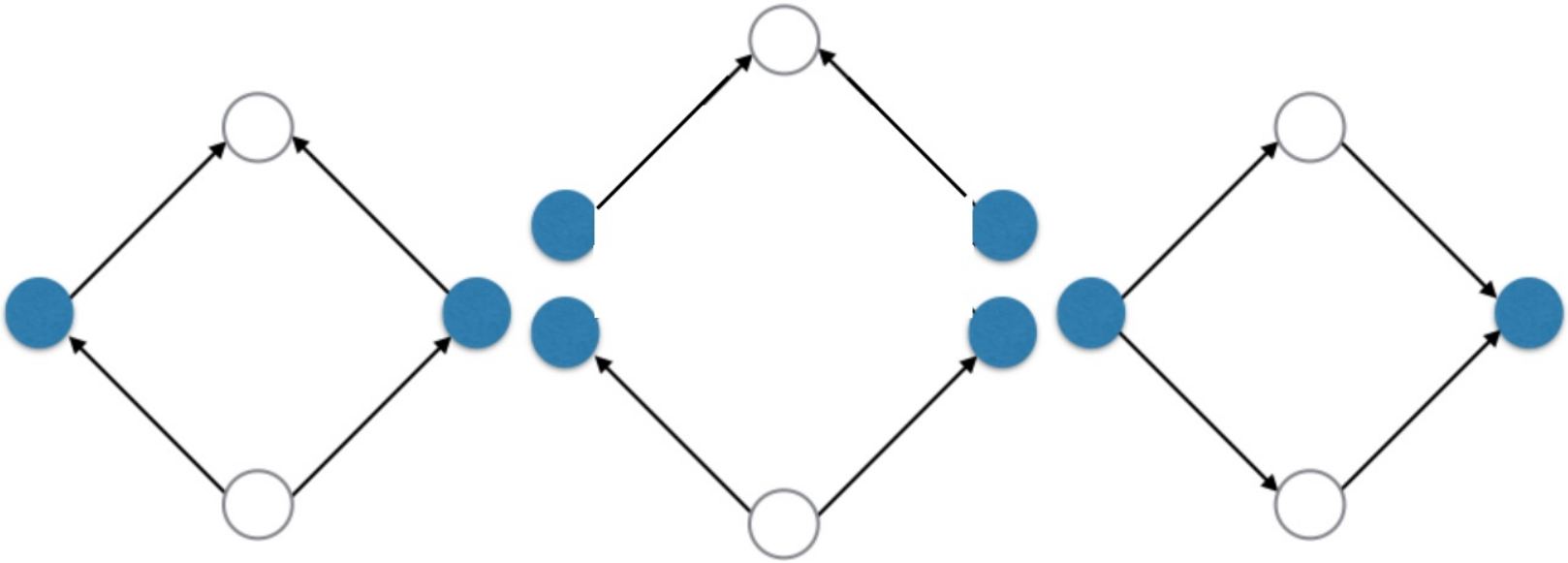
- degré

Conclusion: Voir les noeuds les plus discutés/controversés :

- 18 : Démanteler les bases nucléaires
- 15 : Intervention de l'état
- 8 : Immigration
- 5 : Service national
- 5 : 18000 emplois supprimées
- 5 : Baisse du pouvoir d'achat à court terme
- 4 : Besoin d'un revenu universel pour faire une transition et aider
- 4 : Production électricité par photovoltaique ou eolienne

Algorithmes: Evaluer la qualité d'un argument

- Répartition de poids pour les noeuds:
  - p1 = 3*n, Noeuds attaquent les même noeuds ET sont attaqués par d'autres noeuds
  - p2 = 2*n, Noeuds attaquent OU sont attaqués par les mêmes noeud
  - p3 = 1*n, Relation de transitivité
- P = p1+p2+p3-degré(in)



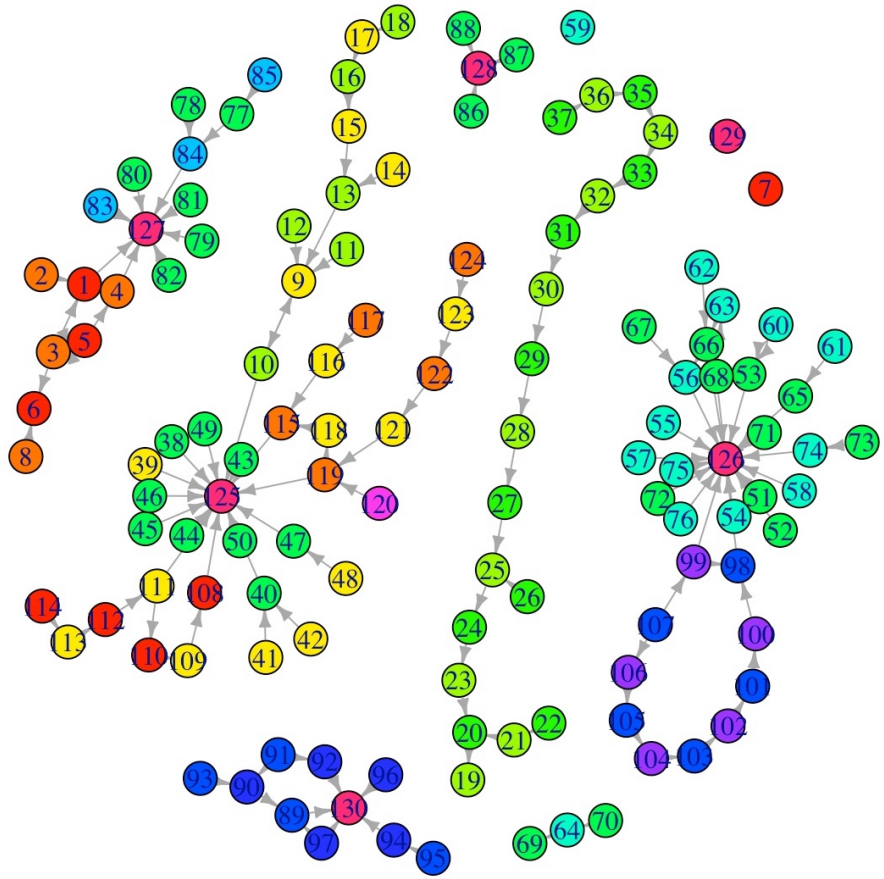
Algorithmes: Résultat

- P:

```
  [1] 32  4 17 19 24  6  0  4  0 37  8  8  6
 [14]  4  3 -1 -1  0 -1 -2  3  0  3 -1 -2  4
 [27]  3 -1 -1 -1 -1 -1 -1 -1 -1 -1  0 30 30
 [40] 28  4  4 30 30 30 30 29  0 30 30 36 36
 [53] 50 39 36 46 36 36  0  8  0  4  9 -1 35
 [66] 42  8 44 -1  0 36 42  0 35 40 40  3  4
 [79] 16 16 16 16 16 14  0  6  6  6 12  6  4
 [92] 13  4  9  0 10 14 38 43  3 -1 -1 -1 -1
[105] -1  3  0 29 -1  3 29 -1 -1  0 28  3  0
[118]  7 28  4  3 -1 -1  0 -3 12 -4 -3  0 -1
```

- Classement:

```
  [1]  53  56  68  99  66  72  75  76  54  98
 [11]  10  51  52  55  57  58  71  65  74   1
 [21]  38  39  43  44  45  46  49  50  47 108
 [31] 111  40 115 119   5   4   3  79  80  81
 [41]  82  83  84  97  92  89 126  96  63  94
 [51]  11  12  60  67 118   6  13  86  87  88
 [61]  90   2   8  14  26  41  42  62  78  91
 [71]  93 120  15  21  23  27  77 100 106 110
 [81] 116 121   7   9  18  22  37  48  59  61
 [91]  70  73  85  95 107 114 117 124 129  16
[101]  17  19  24  28  29  30  31  32  33  34
[111]  35  36  64  69 101 102 103 104 105 109
[121] 112 113 122 123 130  20  25 125 128 127
```



Algorithmes: Conclusion

- Plus un argument est attaqué par des "bons" arguments, plus l'argument est considéré comme mauvais
- Résultat possible:
  - Plus le score est faible, plus l'argument est mauvais
  - Plus le score est élevé, plus l'argument bon

Algorithmes: Personalités similaires

- Utilisation de la fonction similarite:
  - Fillon et Le pen : -10
  - Mélanchon et Le Pen : 3
  - Mélanchon et Hamon : -4
  - Fillon et Hamon : -11
- Interprétation:
  - Résultats surprenants
  - Rapprochement entre extrème droite et gauche forte au niveau de l'argumentation
  - Gauche très divisé (Mélanchon et Hamon)
  - Opposition entre Fillon et Hamon
  - Opposition entre Fillon et Le pen (Droite très variées)
- Insuffisance des données pour une véritable conclusion
- Non prise en compte de la sémantique des données

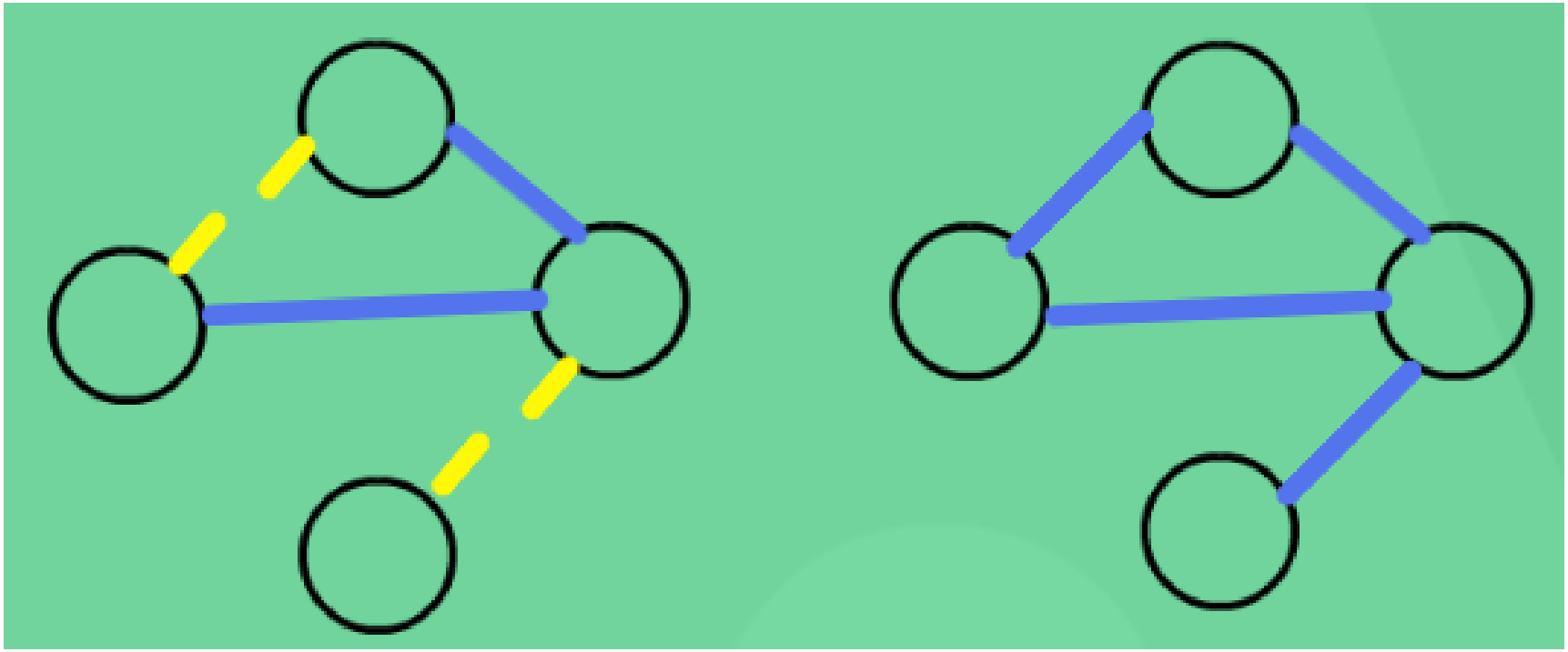
Algorithmes: Prédiction de liens

Prédictions d'existence de lien en fonction de critères/métriques définis:

- Similarité de Jaccard $J(A,B) = \frac{A \cap B}{AUB}$
- Pondération Inverse à la Distance $w_k(x) = \frac{1}{d(x,x_k)^p}$
- Degrès
- Shortest Path

	x	y	jac	AA	AP	SP	classe
0	0	21	0.333333	0.360674	1	2	False
1	0	20	0.333333	0.360674	1	2	False
2	0	19	0.333333	0.360674	2	2	False
3	0	11	0.333333	0.360674	1	3	False
4	0	10	0.333333	0.360674	1	3	False
5	0	9	0.333333	0.360674	1	3	False
6	0	8	0.333333	0.360674	2	3	False
7	0	7	0.333333	0.360674	1	2	False
8	0	6	0.333333	0.360674	1	2	False
9	0	5	0.333333	0.360674	3	2	False
10	0	4	0.333333	0.360674	1	2	False
11	0	3	0.333333	0.360674	1	2	True
12	1	21	0.333333	0.360674	1	2	False
13	1	20	0.333333	0.360674	1	2	False
14	1	19	0.333333	0.360674	2	2	False
15	1	11	0.333333	0.360674	1	3	False
16	1	10	0.333333	0.360674	1	3	False

Algorithmes: Prédiction de liens



Etapes:

- Génération de données
- Création d'un arbre de décision
- Evaluation par validation croisé

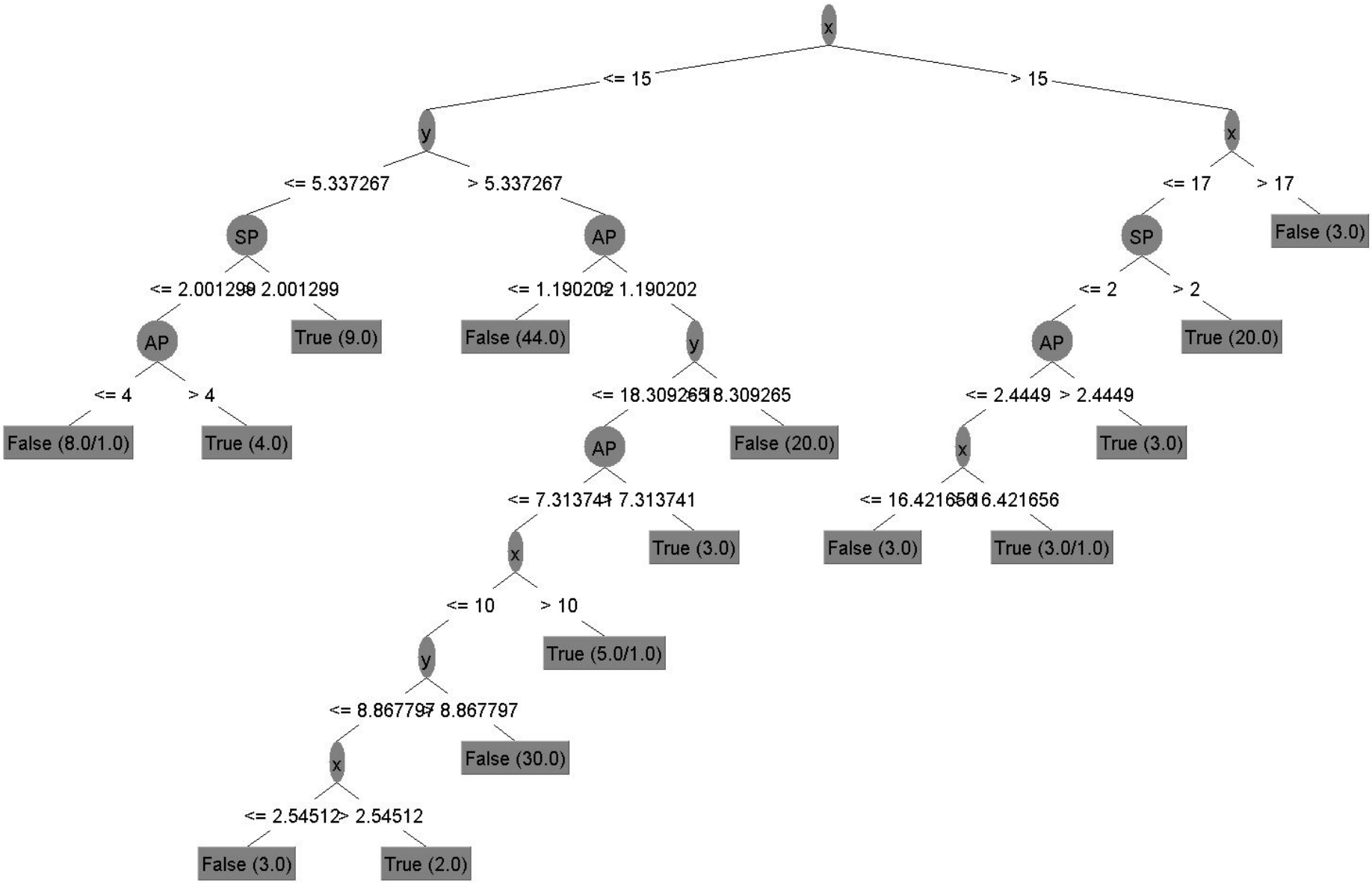
Resultat -1-

```
=== Summary ===

Correctly Classified Instances         112               94.9153 %
Incorrectly Classified Instances         6                5.0847 %
Kappa statistic                          0
Mean absolute error                      0.0766
Root mean squared error                  0.2197
Relative absolute error                 73.4472 %
Root relative squared error             99.5712 %
Total Number of Instances              118

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000     1,000     0,949       1,000   0,974      0,000    0,641     0,963     False
                 0,000     0,000     0,000       0,000   0,000      0,000    0,641     0,078     True
Weighted Avg.    0,949     0,949     0,901       0,949   0,924      0,000    0,641     0,918

=== Confusion Matrix ===

   a   b   <-- classified as
 112   0 |   a = False
   6   0 |   b = True
```

Resultat -2-

```
=== Summary ===

Correctly Classified Instances         139               86.875  %
Incorrectly Classified Instances        21               13.125  %
Kappa statistic                          0.6818
Mean absolute error                      0.1597
Root mean squared error                  0.349
Relative absolute error                 37.9074 %
Root relative squared error             76.1309 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,920    0,250    0,896      0,920    0,907      0,683    0,845     0,888     False
                 0,750    0,080    0,800      0,750    0,774      0,683    0,845     0,761     True
Weighted Avg.    0,869    0,199    0,867      0,869    0,868      0,683    0,845     0,850

=== Confusion Matrix ===

   a   b   <-- classified as
 103   9 |   a = False
  12  36 |   b = True
```

Resultat -2-


x
<= 15
> 15
y
x
<= 5.337267
> 5.337267
<= 17
> 17
SP
AP
SP
False (3.0)
<= 2.001299
> 2.001299
<= 1.190202
> 1.190202
<= 2
> 2
AP
True (9.0)
False (44.0)
y
AP
True (20.0)
<= 4
> 4
<= 18.309265
> 18.309265
<= 2.4449
> 2.4449
False (8.0/1.0)
True (4.0)
AP
False (20.0)
x
True (3.0)
<= 7.313741
> 7.313741
<= 16.421656
> 16.421656
True (3.0)
x
False (3.0)
True (3.0/1.0)
<= 10
> 10
y
True (5.0/1.0)
<= 8.867797
> 8.867797
False (30.0)
x
<= 2.54512
> 2.54512
False (3.0)
True (2.0)

Discussion

- Base de donnée déséquilibré ==> sur-représentation d'une classe
- Métriques non adaptés au problème
- Incapacité à représenter la sémantique du problème
- Problème mal abordé

Conclusion

- Travail axé sur le fond via les graphes
- Retranscription des liens entre arguments
- Analyse des interactions entre arguments (similarité arguments, sujets les plus discutés, founis, ...)
- prédiction de lien

Perspectives

- Analyse du discours
  - Extraction de sens
  - Modélisation de celui ci dans un graphe

Perspectives

Identifier une personne à un bord politique d'après son argumentation (forme et non fond)
Abordé le problème par une analyse de texte:

- L'utilisation de phrases longue
- Ratio champs lexicaux d'une idée / tout les mots
- Utilisation du vouvoiement/tutoiement
- Voir si le discours est policé ou s'il utilise des mots de tendance "cru" (noir, ...)
- Utilisation de figures de styles
- Reformulation de la question
- Mesure de tempérament
- Esquive d'une question
- Historique

Bibliographie

Pour la génération de données:

- vidéo de débats politique de 'L Emission Politique' sur youtube.com ou sur francetvinfo.fr/replay-magazine/france-2/l-emission-politique
- Les programmes politiques sur .lemonde.fr/personnalite/nom de la personne/programme
- Apache beam : https://beam.apache.org/
- Neo4j : https://neo4j.com/developer/

Merci de votre attention
Avez vous des questions ?