Algorithmique et Structuration de données (TALA330A L3 INALCO) -

Algoritihmes : Distances et Similarités pour le TAL

Caroline Koudoro-Parfait

Crédits : Gaël Lejeune

caroline.parfait@sorbonne-universite.fr

Pourquoi la similarité ?

Pourquoi la similarité?

Une opération essentielle : comparer

- Les opérateurs informatiques (=, <, > ...) sont puissants

Pourquoi la similarité?

Une opération essentielle : comparer

- Les opérateurs informatiques (=, <, > ...) sont puissants
- Par contre on ne sait pas bien faire '

Pourquoi la similarité?

Une opération essentielle : comparer

- Les opérateurs informatiques (=, <, > ...) sont puissants
- Par contre on ne sait pas bien faire '

Pourquoi ?

Pourquoi la similarité?

Une opération essentielle : comparer

- Les opérateurs informatiques (=, <, > ...) sont puissants
- Par contre on ne sait pas bien faire '

Pourquoi ?

- ”pareil” (=) → facile à définir
- ”plus petit” dépend de ce qu’on observe : chaînes, entiers, dates. . .
- ”**presque**” (pareil) c’est plus compliqué

Pourquoi la similarité?

Une opération essentielle : comparer

- Les opérateurs informatiques (=, <, > ...) sont puissants
- Par contre on ne sait pas bien faire '

Pourquoi ?

- ”pareil” (=) → facile à définir
- ”plus petit” dépend de ce qu’on observe : chaînes, entiers, dates. . .
- ”**presque**” (pareil) c’est plus compliqué

Figure 1 – ”Une bière ca reste une bière”, XKCD Isn’t Funny http://xkcdisntfunny.blogspot.com/

Veuillez définir similaire ?

Qui est de même nature (TLFI)

- le RER et le métro c’est pareil, c’est sur des rails

Veuillez définir similaire ?

Qui est de même nature (TLFI)

- le RER et le métro c’est pareil, c’est sur des rails
- Lego masters c’est comme Koh lanta

Veuillez définir similaire ?

Qui est de même nature (TLFI)

- le RER et le métro c’est pareil, c’est sur des rails
- Lego masters c’est comme Koh lanta
- C’est un goéland ou une mouette ?
  Facile, le goéland est plus grand

Veuillez définir similaire ?

Qui est de même nature (TLFI)

- le RER et le métro c’est pareil, c’est sur des rails
- Lego masters c’est comme Koh lanta
- C’est un goéland ou une mouette ?
  Facile, le goéland est plus grand
- Le train est-il plus proche de l’avion ou du métro

Similarité/Equivalence VS Identité/Egalité

- etat VS état
- EgalVS égal
- Björn VS Bjorn

Veuillez définir similaire ?

Qui est de même nature (TLFI)

- le RER et le métro c’est pareil, c’est sur des rails
- Lego masters c’est comme Koh lanta
- C’est un goéland ou une mouette ?
  Facile, le goéland est plus grand
- Le train est-il plus proche de l’avion ou du métro

Similarité/Equivalence VS Identité/Egalité

- etat VS état
- EgalVS égal
- Björn VS Bjorn

Comment définir la nature d’un objet ?

→ on décompose, on regarde des traits, on type

Exemple concret sur des chaînes (I)

Comment encoder une intuition sur la ressemblance ?

Comparer avec des opérateurs de base (=, <, >, *IN* . . .) :

str 1 : ”Sur le climat, il n’y a pas de plan B. Car il n’y a pas de planète B”[1]

str 2 : ”Il n’y a pas de plan B car il n’y a pas de planète B”[2]

str 3 : ”Sur la planète Mars il n’y a pas de plan B pour le climat”

- str1 = str2 tout comme str2 = str3
- str1 NOT IN str2 ET str2 NOT IN str1

1. E.Macron 2017
2. Ban Ki Moon 2016

Exemple concret sur des chaînes (I)

Comment encoder une intuition sur la ressemblance ?

Comparer avec des opérateurs de base (=, <, >, *IN* . . .) :

str 1 : "Sur le climat, il n'y a pas de plan B. Car il n'y a pas de planète B"

str 2 : "Il n'y a pas de plan B car il n'y a pas de planète B"[2]

str 3 : "Sur la planète Mars il n'y a pas de plan B pour le climat"

- str1 = str2 tout comme str2 = str3
- str1 NOT IN str2 ET str2 NOT IN str1

Tokens en commun

str1/str2	8	B	_	a	de	il	_	n'y	pas	plan	planète
str1/str3	10	B	Sur	a	de	il	le	n'y	pas	plan	planète
str2/str3	8	B	_	a	de	il	_	n'y	pas	plan	planète

1. E.Macron 2017
2. Ban Ki Moon 2016

Exemple concret sur des chaînes (II)

str 1 : ”Sur le climat, il n’y a pas de plan B. Car il n’y a pas de planète B”

str 2 : ”Il n’y a pas de plan B car il n’y a pas de planète B”

str 3 : ”Sur la planète Mars il n’y a pas de plan B pour le climat”

Paire	Dist	len(LCS)	LCS
str1/str2	18	28	”il n’y a pas de planète B”
str1/str3	35	23	”il n’y a pas de plan B”
str2/str3	33	22	”l n’y a pas de plan B”

Table 1 – Distance de Levenshtein en caractères et Longest common substring (LCS)

Important :

- Manière dé définir les objets/leurs caractéristiques
- Potentiel de la redescription (lemmatisation, correction . . .)
- Choix de la mesure de similarité/distance

Définition et et cas d’utilisations

Définitions

Similarité (ou affinité en clustering/recommandation) :
mesure de la proximité de deux objets

Distance éloignement de deux objets
Objets définis par des traits, des caractéristiques

Définitions

Similarité (ou affinité en clustering/recommandation) : mesure de la proximité de deux objets

Distance éloignement de deux objets

Objets définis par des traits, des caractéristiques

Sim VS Dist quand Sim est défini entre 0 et 1 : Sim = 1 - Dist

Différents types d’objets :

Définitions

Similarité (ou affinité en clustering/recommandation) :
mesure de la proximité de deux objets

Distance éloignement de deux objets
Objets définis par des traits, des caractéristiques

Sim VS Dist quand Sim est défini entre 0 et 1 : Sim = 1 - Dist

Différents types d'objets :

- chaînes
- séquences ADN
- documents
- données météo . . .

Définitions

Similarité (ou affinité en clustering/recommandation) :
mesure de la proximité de deux objets

Distance éloignement de deux objets
Objets définis par des traits, des caractéristiques

Sim VS Dist quand Sim est défini entre 0 et 1 : Sim = 1 - Dist

Différents types d'objets :

- chaînes
- séquences ADN
- documents
- données météo . . .

Rapprocher des objets/instances c'est utile :

Définitions

Similarité (ou affinité en clustering/recommandation) :
mesure de la proximité de deux objets

Distance éloignement de deux objets
Objets définis par des traits, des caractéristiques

Sim VS Dist quand Sim est défini entre 0 et 1 : Sim = 1 - Dist

Différents types d'objets :

- chaînes
- séquences ADN
- documents
- données météo . . .

Rapprocher des objets/instances c'est utile :

- classification
- auto-complétion
- recommandation

Différentes utilisations

Que mesure-t-on ?

- Distance kilométrique/temps/coût/danger . . .
- Distance sémantique, lexicographique . . .

Différentes utilisations

Que mesure-t-on ?

- Distance kilométrique/temps/coût/danger . . .
- Distance sémantique, lexicographique . . .

Dans quoi ?

- Des séquences, des graphes . . .

Comment ?

- Contexte : importance ou pas de l’ordre (par ex. présence en classe)
- Pondération : toutes les différences se valent-elles (par ex. absence d’un joueur dans une équipe sportive)

Différentes utilisations

Que mesure-t-on ?

- Distance kilométrique/temps/coût/danger . . .
- Distance sémantique, lexicographique . . .

Dans quoi ?

- Des séquences, des graphes . . .

Comment ?

- Contexte : importance ou pas de l'ordre (par ex. présence en classe)
- Pondération : toutes les différences se valent-elles (par ex. absence d'un joueur dans une équipe sportive)
- Les deux ? (par ex. absence d'une séquence importante de deux éléments qui ne le sont pas individuellement

Types de Données et propriétés

Propriétés de la similarité

- Avec des données séquentielles, plus facile de mesurer l’identité stricte (tri VS sac)

Types de Données et propriétés

Propriétés de la similarité

- Avec des données séquentielles, plus facile de mesurer l’identité stricte (tri VS sac)
- Avec des données non séquentielles (ex BOW), plus facile de mesurer une distance

Types de Données et propriétés

Propriétés de la similarité

- Avec des données séquentielles, plus facile de mesurer l’identité stricte (tri VS sac)
- Avec des données non séquentielles (ex BOW), plus facile de mesurer une distance
- L’égalité est transitive, la similarité pas nécessairement

Types de Données et propriétés

Propriétés de la similarité

- Avec des données séquentielles, plus facile de mesurer l’identité stricte (tri VS sac)
- Avec des données non séquentielles (ex BOW), plus facile de mesurer une distance
- L’égalité est transitive, la similarité pas nécessairement

Algorithmes

Sur des données purement séquentielles

- Distance de Levenshtein
- Word Error rate(WER), Character Error Rate (CER) . . .
- Longest Common Substring
- Sous séquences fermées fréquentes (ou répétées maximales)

Algorithmes

Sur des données purement séquentielles

- Distance de Levenshtein
- Word Error rate(WER), Character Error Rate (CER) . . .
- Longest Common Substring
- Sous séquences fermées fréquentes (ou répétées maximales)

Sur des données où l’aspect séquentiel n’est pas pertinent

- Indice de Jaccard
- Distance Euclidienne
- Distance Cosinus . . .

Algorithmes

Sur des données purement séquentielles

- Distance de Levenshtein
- Word Error rate(WER), Character Error Rate (CER) . . .
- Longest Common Substring
- Sous séquences fermées fréquentes (ou répétées maximales)

Sur des données où l'aspect séquentiel n'est pas pertinent

- Indice de Jaccard
- Distance Euclidienne
- Distance Cosinus . . .

Beaucoup d'autres cas

- Données structurées ?
- Graphes, arbres . . .
- Texte + mise en forme, Texte + structure . . .