Levenshtein et Classification

Sylvain Gault

8 septembre 2024

Mesurer la similarité

Correcteur orthographique

- L'utilisateur a tapé « *graffe* », quel est le mot le plus proche ?
  - gaffe
  - greffe
  - ...
  - graphe

Bioinformatique

- Aligner deux séquences de nucléotides avec des mutations
  - `-AGGCTATCACCT`
  - `TAG-CTATCGCCT`

Autres

- Traduction automatique, Extraction d'informations, Reconnaissance de la parole, etc.

Distance d'édition

Définition : distance d'édition entre deux chaînes de caractères

- Nombre minimum d'opérations d'édition
  - Insertion
  - Suppression
  - Substitution
- Nécéssaires pour transformer l'une en l'autre

Autres opérations possibles, mais rarement utilisés

- Déplacement longue distance
- Transposition
- ...

Distance d'édition minimum

Exemple

- `INTE-NTION`
- `-EXECUTION`
- `dss-is----`

Calcul

- Distance 5 si toutes les opérations ont un coût 1
- Distance 8 si les substitutions ont un coût 2 (Lenvenshtein)

Autres exemples

Traduction automatique

- Mesurer la qualité de la traduction
  - H : Spokesman confirms senior gorvernment adviser was shot
  - M : Spokesman said senior adviser was shot dead
- 1 substitution, 2 insertions, 1 suppression

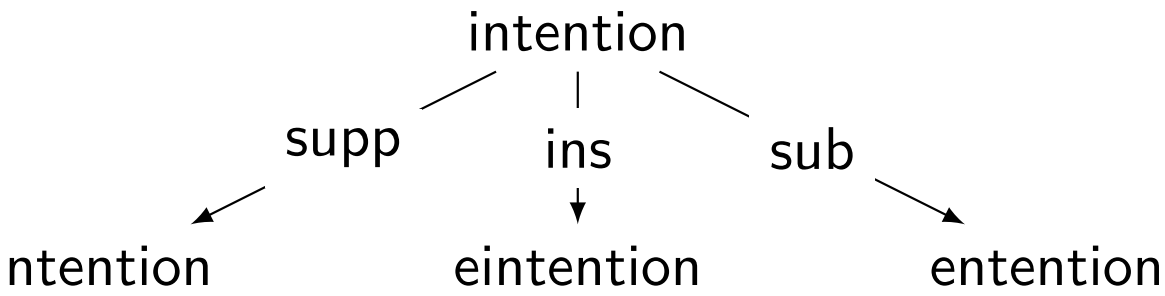
Extraction d'Entités Nommées

- Détecter quand il s'agir de la même entité
  - ***IBM Inc.*** *announced today*
  - ***IBM*** *profits*
  - ***Stanford President John Hennessy*** *announced yesterday*
  - *for* ***Stanford University President John Hennessy***
- Peu de différence : Même entité

Trouver la distance d'édition minimum

Technique

- Recherche un chemin (séquence d'éditions) depuis le début jusqu'à la fin
- État initial : Le mot à transformer
- État final : Le mot à recréer
- Opérateurs : Insertion, suppression, substitution
- Coût : Ce qu'on veut minimiser : Le nombre d'éditions



Recherche d'édition minimum

Problèmes et solutions

- L'espace des séquences d'édition est énorme
- Besoin d'une stratégie intelligente

Recherche d'édition minimum

Problèmes et solutions

- L'espace des séquences d'édition est énorme
- Besoin d'une stratégie intelligente
- Beaucoup de chemins amènent au même état

Recherche d'édition minimum

Problèmes et solutions

- L'espace des séquences d'édition est énorme
- Besoin d'une stratégie intelligente
- Beaucoup de chemins amènent au même état
- On a besoin uniquement du plus court chemin

Définition de la distance d'édition minimum

Formalisme

- Chaîne X de longueur n
- Chaîne Y de longueur m
- $D(i, j)$ la distance d'édition minimum entre $X[1..i]$ et $Y[1..j]$
- On recherche $D(n, m)$

Programmation Dynamique pour distance d'édition minimum

Programmation dynamique

- Méthode récursive avec cache
- Résoudre un problème par combiaison de solutions à des sous-problèmes
- Approche Bottom-up
  - Calcul de $D(i, j)$ pour des petites valeurs de i, j
  - En déduire la valeur de $D(i, j)$ pour des i, j plus grands
  - $\rightarrow$ Calculer $D(i, j)$ pour tous les $i \leq n$ et tous les $j \leq m$

Distance de Levenshtein

Initialisation

- $D(i, 0) = i$ i suppressions
- $D(0, j) = j$ j insertion

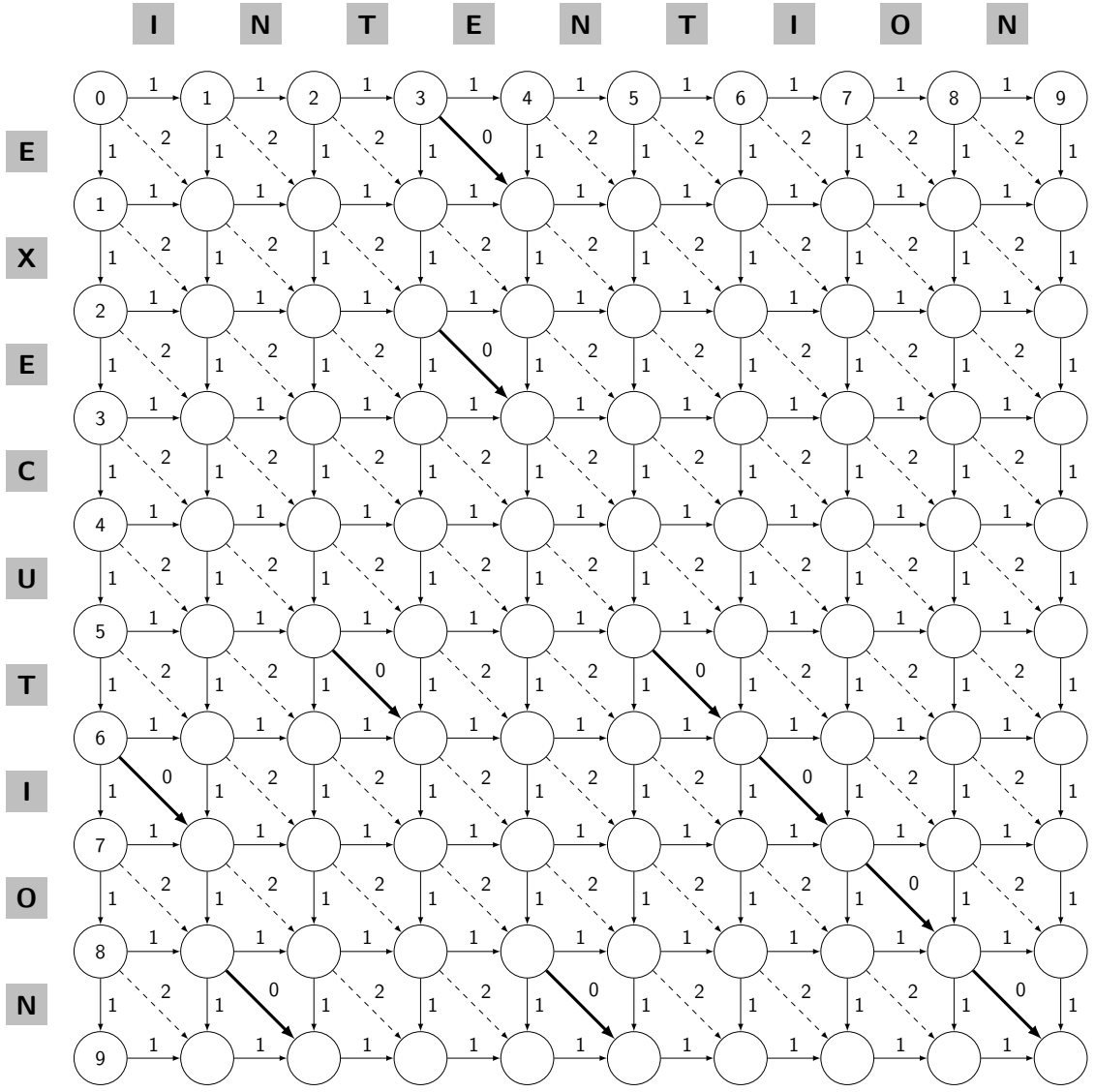
Relation de récurrence

- $D(i, j) =$ minimum de
  - $D(i-1, j) + 1$ suppression
  - $D(i, j-1) + 1$ insertion
  - $D(i-1, j-1) + 2$ Si $X[i] \neq Y[j]$ substitution
  - $D(i-1, j-1) + 0$ Si $X[i] = Y[j]$ matching

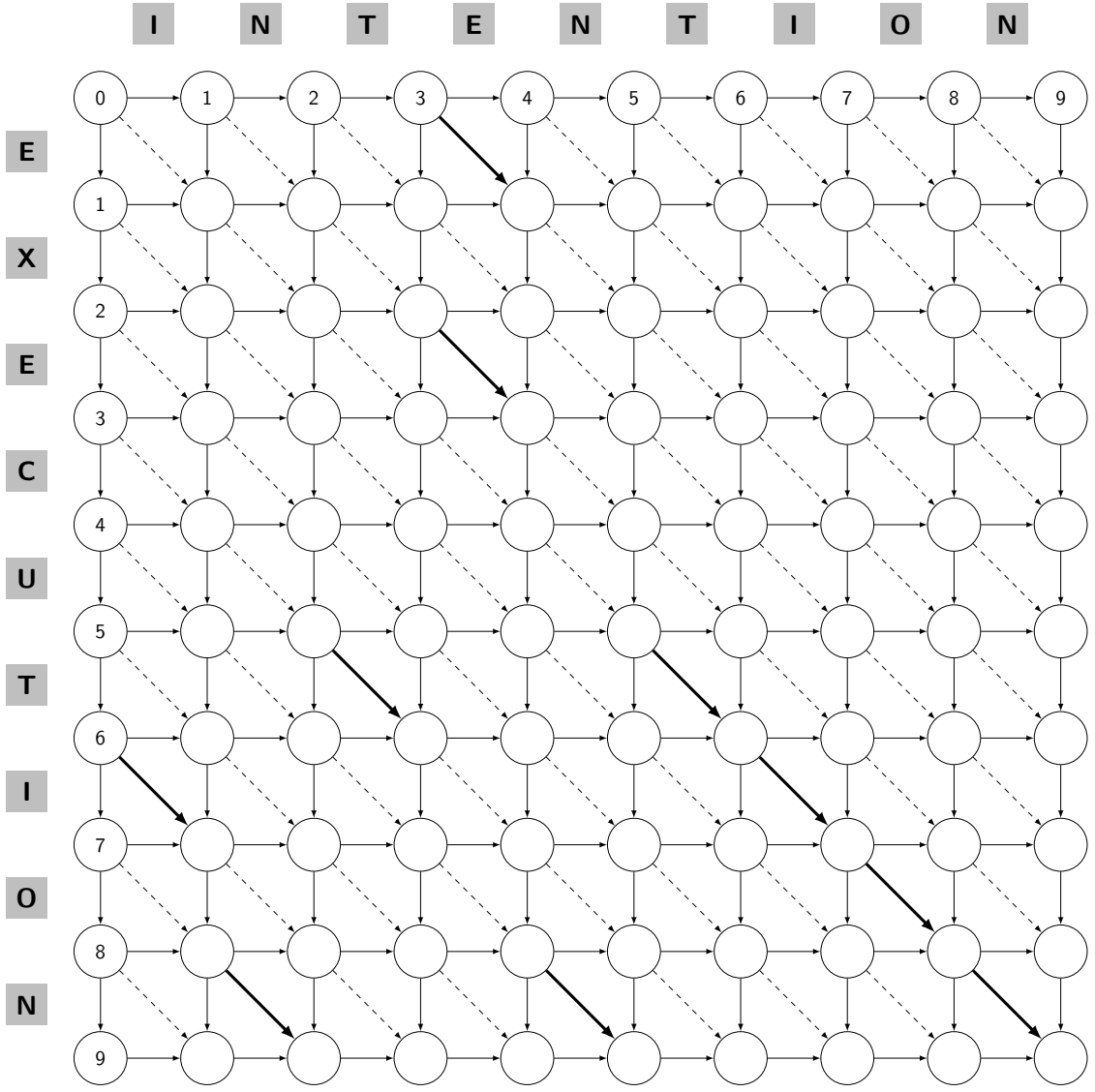
Terminaison

- Quand $D(n, m)$ est connu

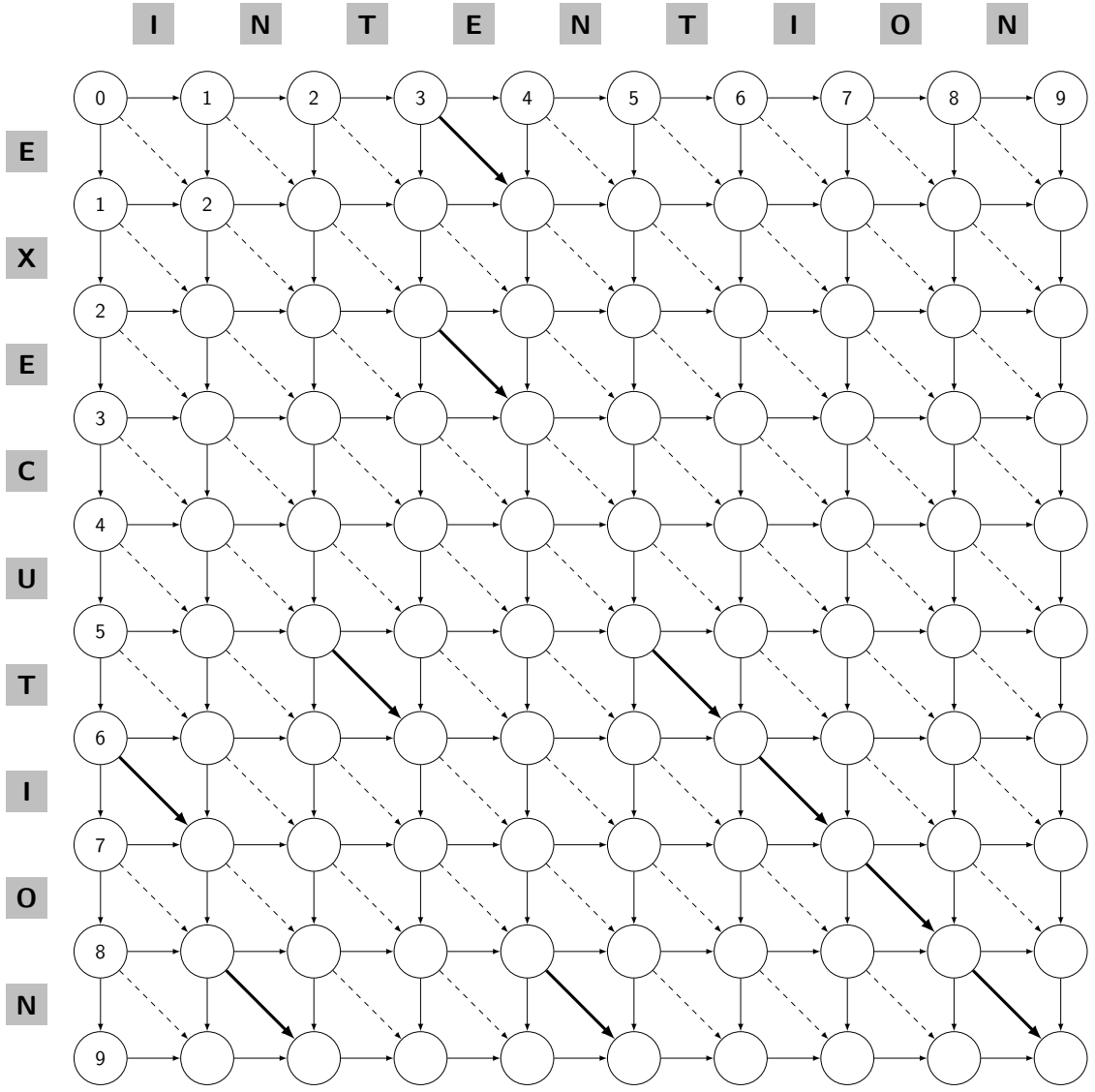
Distance d'édition



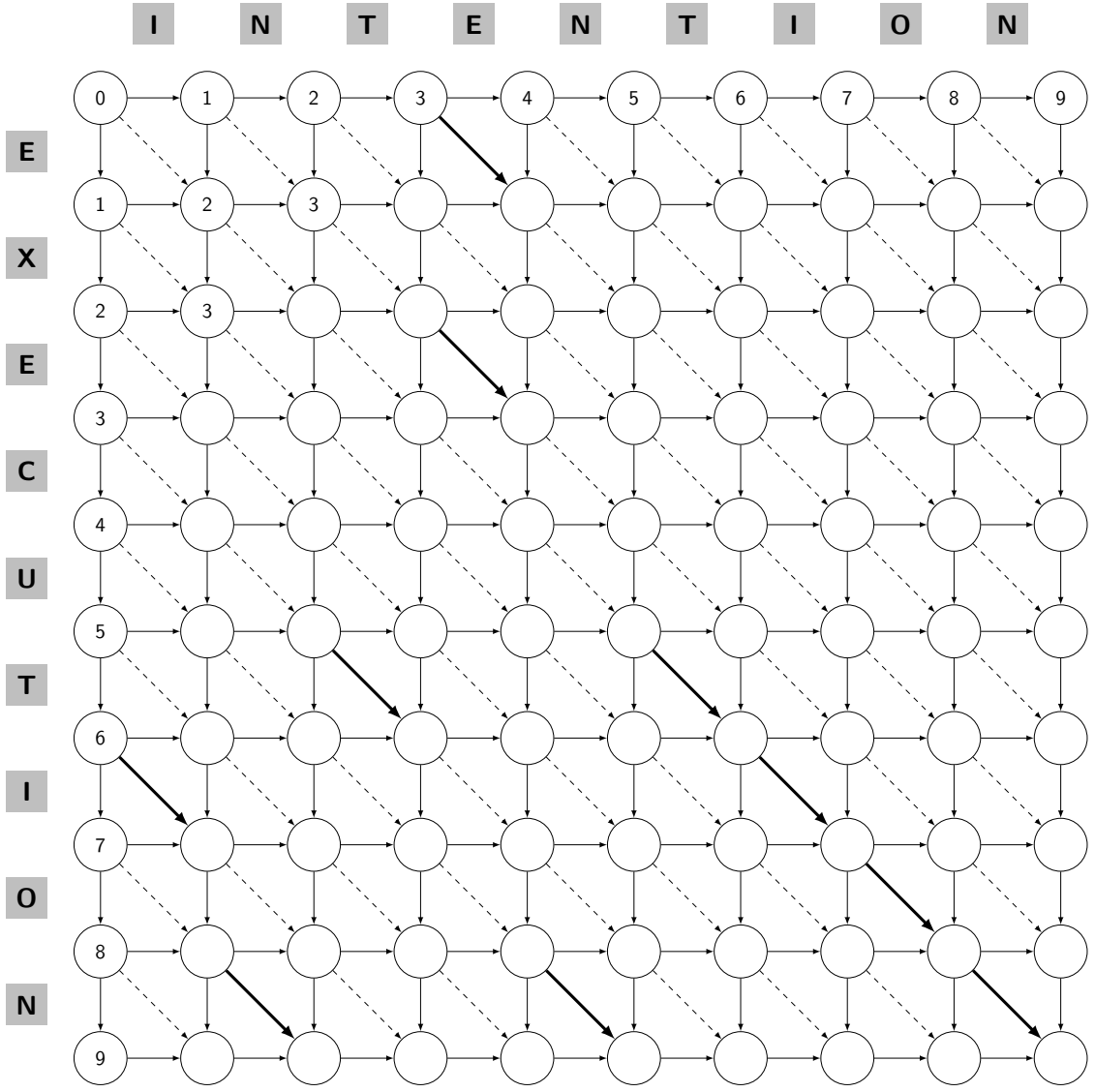
Distance d'édition



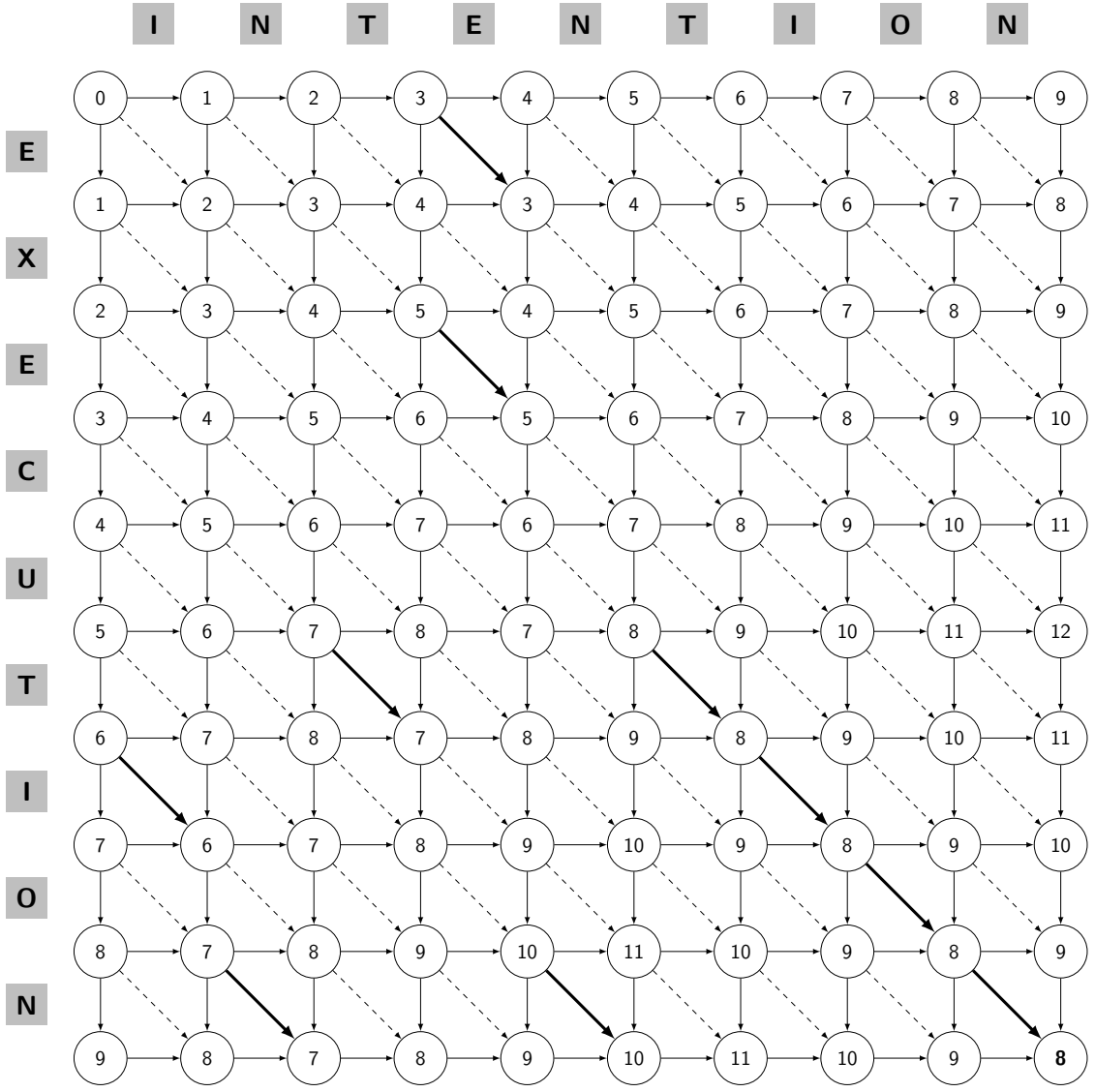
Distance d'édition



Distance d'édition



Distance d'édition



Trace du chemin

Backtrack

- En pratique, il faut garder une trace du chemin vers l'arrivée
- À chaque nœud, garder en mémoire le ou les nœuds précédents de coût minimum
- Ce chemin donne la séquence d'insertions, suppressions, substitutions, correspondances

Complexité

En temps

- $\mathcal{O}(nm)$
- Il faut remplir toute la grille

En espace

- $\mathcal{O}(nm)$
- Il faut remplir toute la grille

De la backtrace

- $\mathcal{O}(n+m)$
- Au pire, on a que des suppression et que des insertions

Distance d'édition pondérée

Podération

- Toutes les erreurs ne sont pas aussi fréquentes
  - Dépend du type d'erreurs considérés : fautes de frappe sur un clavier azerty ? Clavier virtuel de smartphone ? Fonction "swipe" ? Erreurs de reconnaissance vocale ? etc.
- Coût différents pour l'insertion de différentes lettres
- Coût différent pour la suppression de différentes lettres
- Coût différent pour la substitutions de différentes paires de lettres
- Algorithme autrement identique

Questions ?

Questions ?

- Questions ?

TP

TP

- TP

Exemple

Spam ou non ?

Our Names are Frances and Patrick Connolly and our foundation is donating (£1.5 Million Pounds) to you. Contact us via my email at (capinolly@gmail.com) for further details.
Best Regards,
Frances & Patrick Connolly,

Identification d'auteurs

Comment savoir qui a écrit quel texte ?

- En 1787-1788, des lettres anonymes ont été envoyé à l'état de New-York pour ratifier la constitution
- Leurs 3 auteurs étaient connus
- L'auteur de 12 de ces lettres restait non-identifié
- Résolu en 1963 Mosteller et Wallace avec une méthode Bayesienne

Identification du genre de l'auteur

Auteur masculin ou féminin ?

- Nombre de pronoms, de déterminants, de groupe nominaux, etc.
- Féminin : Plus de pronoms
- Masculin : Plus de faits et déterminants

Analyse de sentiments

Avis sur les films : Positif ou négatif ?

- Incroyablement décevant

Analyse de sentiments

Avis sur les films : Positif ou négatif ?

- Incroyablement décevant
- Rempli de personnages loufoques, de savante satire et d'énormes retournements de situation

Analyse de sentiments

Avis sur les films : Positif ou négatif ?

- Incroyablement décevant
- Rempli de personnages loufoques, de savante satire et d'énormes retournements de situation
- Ce film est le plus grand échec jamais filmé

Analyse de sentiments

Avis sur les films : Positif ou négatif ?

- Incroyablement décevant
- Rempli de personnages loufoques, de savante satire et d'énormes retournements de situation
- Ce film est le plus grand échec jamais filmé
- Pathétique. Le pire était le passage avec les scènes de boxe

Catégorisation d'articles scientifiques

Exemples

- Antagonistes et inhibiteurs
- Chimie
- Médication
- Embryologie
- Épidémiologie
- ...

Classification de texte

Type de tâches

- Assigner une catégorie de sujet, un genre, ...
- Détection de spam
- Identification d'auteurs
- Identification d'âge et genre
- Analyse de sentiment
- Identification de langue
- ...

Classification de texte

Définition

- Entrée :
  - Un document d
  - Un ensemble déterminé de classes $C = c_1, c_2, ..., c_k$
- Sortie :
  - Une classe $c \in C$

Méthodes de classification

Méthodes manuelles

- Règles basées sur des combinaisons mots ou autres *features*
  - « dollar » + « donated »
- Fonctionne raisonnablement bien
- Maintenir les règles à jour demande beaucoup d'effort
- $\rightarrow$ Utile comme premier filtre

Méthodes de classification

Machine Learning Supervisé

- Entrée :
  - Un document d
  - Un ensemble déterminé de classes $C = c_1, c_2, ..., c_k$
  - Un jeu de données d'entraînement de n exemples correctement classifiés $(d_1, c_1), (d_2, c_2), ..., (d_n, c_n)$
- Sortie :
  - Un classifieur $\gamma : d \rightarrow c$

Méthodes de classification

Méthodes existantes

- Classifieur bayesien naïf
- Régression logistique
- SVM (Support-Vector Machine)
- KNN (k-Nearest Neighbors)
- ...

Méthodes de classification

Méthodes existantes

- **Classifieur bayesien naïf**
- Régression logistique
- SVM (Support-Vector Machine)
- KNN (k-Nearest Neighbors)
- ...

Méthodes de classification

Méthodes existantes

- **Classifieur bayesien naïf**
- **Régression logistique**
- SVM (Support-Vector Machine)
- KNN (k-Nearest Neighbors)
- ...

Intuition de la classification bayesienne naïve

Présentation

- Basée sur le théorème de Bayes
- Représente les documents comme des sacs de mots (Bag of Words)

Bag of Words

Description

- Perd l'ordre des mots (comme un ensemble)
- Garde l'information du nombre d'occurence des mots
- Peut être appliqué sur un sous-ensemble des mots

Exemple : entrée

- « je me présente, je m'appelle Henry »

Exemple : sortie

- « je » : 2
- « me » : 1
- « présente » : 1
- ...

Bag of Words

Utilisation

- Affecter un certain poids à chaque mot pour chaque classe
- Ex. : « génial » vs. « nul », « algorithm » vs. « protein »

Théorème de Bayes appliqué aux documents

Théorème de Bayes

- Pour un document d et une classe c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Utilisation du théorème de Bayes

Utilisation

$$\begin{aligned} c^* &= \underset{c \in C}{\text{argmax}}\, P(c|d) \\ &= \underset{c \in C}{\text{argmax}}\, \frac{P(d|c)P(c)}{P(d)} \\ &= \underset{c \in C}{\text{argmax}}\, P(d|c)P(c) \end{aligned}$$

- Produit de la vraisemblance et de la probabilité à priori de la classe c

Classifieur bayesien naïf

Calcul

$$c^* = \underset{c \in C}{\text{argmax}}\, P(d|c)P(c)$$
$$= \underset{c \in C}{\text{argmax}}\, P(x_1, x_2, ..., x_n|c)P(c)$$

- x_i les *features* du documents : liste des mots, longueur, etc.

Classifieur bayesien naïf

Calcul

$$c^* = \underset{c \in C}{\text{argmax}}\, P(x_1, x_2, ..., x_n|c)P(c)$$

- $P(c)$: Probabilité d'une classe c quelque soit le document
  - Probabilité d'un spam quelque soit le mail
  - Probabilité d'un avis négatif quelque soit le commentaire
  - Probabilité d'une autrice quelque soit l'œuvre

Classifieur bayesien naïf

Calcul

$$c^* = \underset{c \in C}{\text{argmax}}\, P(x_1, x_2, ..., x_n|c)P(c)$$

- $P(c)$: Probabilité d'une classe c quelque soit le document
  - Probabilité d'un spam quelque soit le mail
  - Probabilité d'un avis négatif quelque soit le commentaire
  - Probabilité d'une autrice quelque soit l'œuvre
- Fréquence d'apparition de la classe c dans le corpus d'entraînement

Classifieur bayesien naïf

Calcul

$$c^* = \underset{c \in C}{\text{argmax}}\, P(x_1, x_2, ..., x_n|c)P(c)$$

- Calculer directement $P(x_1, x_2, ..., x_n|c)$ nécessiterait d'optimiser $\mathcal{O}(|X|^n|C|)$ variables
- Seulement possible avec énormément d'exemples

Supposition d'indépendance : Naïveté

But

- Calculer $P(x_1, x_2, ..., x_n|c)$

Suppositions

- **Bag of Words** : Supposition que la position des mots n'a pas d'importance
- **Indépendance conditionnelle** : Suppose que la probabilité des $P(x_i|c_j)$ sont indépendantes pour une classe c donnée

Résultat

$$P(x_1, x_2, ..., x_n|c) = P(x_1|c)P(x_2|c)...P(x_n|c)$$

Classifieur bayesien naïf multinomial

Formule non-naïve

$$c^* = \underset{c \in C}{\text{argmax}}\, P(x_1, x_2, ..., x_n|c)P(c)$$

Formule naïve

$$c^* = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{x \in X} P(x|c)$$

Application à la classification de texte

Application

$$c^* = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{x \in X} P(x|c)$$

- Pour chaque classe c_j
- Calculer $P(c_j) \prod_{w_i \in BoW} P(w_i|c_j)$
- Retenir la classe qui a la probabilité la plus élevée

Calcul du produit en pratique

Problème

$$c^* = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{x \in X} P(x|c)$$

- Les valeurs de $P(X|c)$ sont petites (très inférieures à 1)
- Leur produit est encore plus petit
- On arrive à 0 et plus rien ne se passe

Calcul du produit en pratique

Solution

- Passer au log

$$c^* = \underset{c \in C}{\text{argmax}}\, log\left(P(c)\prod_{x \in X} P(x|c)\right)$$
$$= \underset{c \in C}{\text{argmax}}\, log(P(c)) + \sum_{x \in X} log(P(x|c))$$

- Le log change les produits en sommes
- Mais ne change pas l'ordre des classes
  - $0 < a < b \Leftrightarrow log(a) < log(b)$
- C'est un max sur une somme de poids
- Le classifieur Bayesien naïf est un classifeur linéaire

Application à la classification de texte

Application

$$c^* = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{w_i \in BoW} P(w_i|c_j)$$

- Comment calculer ces $P(w_i|c_j)$?

Entraînement d'un modèle bayesien naïf

Première idée

- Utiliser les fréquences d'apparition des mots dans les données d'entraînement
- $P(c_j)$: Probabilité d'une classe : pourcentage de documents de cette classe dans le corpus d'entraînement
  - Pourcentage de spam, pourcentage de femmes, ...
- $P(w_i|c_j)$: Parmi tous les documents de classe c_j, quelle est la probabilité de rencontrer le mot w_i ?

Entraînement d'un modèle bayesien naïf

Première idée

- Utiliser les fréquences d'apparition des mots dans les données d'entraînement
- $P(c_j)$: Probabilité d'une classe : pourcentage de documents de cette classe dans le corpus d'entraînement
  - Pourcentage de spam, pourcentage de femmes, ...
- $P(w_i|c_j)$: Parmi tous les documents de classe c_j, quelle est la probabilité de rencontrer le mot w_i ?
  - Sélectionner tous les documents de classe c_j
  - Compter le pourcentage d'occurrence du mot w_i par rapport à tous les mots.

Entraînement d'un modèle bayesien naïf

Problème

- $\widehat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$
- Si un mot n'a pas été vu dans l'ensemble d'entraînement
- $\widehat{P}(\text{"}fantastique\text{"}|positif) = 0$
- Donc la classe *positif* ne sera jamais sélectionnée avec la formule :
- $c^* = \underset{c_j \in C}{\text{argmax}}\, P(c_j) \prod_{w_i \in BoW} P(w_i|c_j)$

Entraînement d'un modèle bayesien naïf

Solution

- Lissage Laplacien (Lissage additif)

$$\widehat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V} (count(w, c_j) + 1)}$$

$$= \frac{count(w_i, c_j) + 1}{\left(\sum_{w \in V} count(w, c_j) \right) + |V|}$$

Entraînement

Étapes

- À partir du corpus d'entraînement, extraire le *Volcabulaire* V
- Calculer les $P(c_j)$ comme le pourcentage de documents de la classe c_j parmi tout le corpus.
- Calculer les $P(w_i|c_j)$
  - Concaténer tous les documents de la classe c_j dans un méga-document $Text_j$
  - Pour chaque mot $w_k \in V$:
    - $n_k \leftarrow$ nombre d'occurrence du mot w_k dans $Text_j$
    - $P(w_i|c_j) \leftarrow \frac{n_k+1}{n+|V|}$

Optimisation pour l'analyse de sentiments

Fréquence des mots

- L'apparition des mots est plus importante que leur fréquence
  - Le fait que le mot *fantastique* apparaîsse nous apporte beaucoup d'information
  - Le fait qu'il apparaîsse 5 fois, beaucoup moins

Classifieur Bayesien Naïf Binaire

- Ou binary NB
- Coupe le comptage des mots à 1
- Note : différent du classifieur Bayesien Naïf de Bernoulli

Questions ?

Questions ?

- Questions ?

TP

TP

- TP

Régression logistique

Généralités

- Outil important dans beaucoup de sciences naturelles et sociales
- Modèle de base pour ML
- Base des réseaux de neurones (1 ReLU = stack de sigmoïdes)

Modèles génératifs et discriminatifs

Classifieur Bayesien Naïf

- Génératif
- Il construit une représentation explicite des classes
  - Quels mots sont probablement du spam
  - Quels features d'une image sont très *cat-like* ou *dog-like*
- $c^* = \text{argmax}_{c \in C} P(d|c)P(c)$

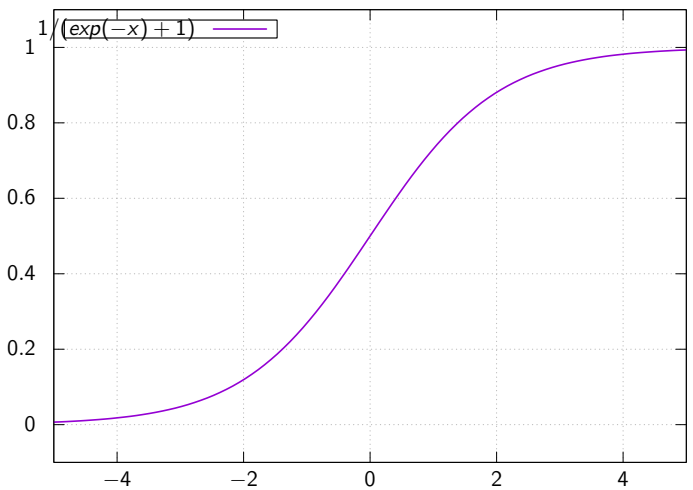
Régression logistique

- Discriminatif
- Il essaye juste de distinguer les classes sans apriori sur ce qui est pertinent
  - Les spams contiennent le symbole \$
  - Les chiens ont des colliers
- $c^* = \text{argmax}_{c \in C} P(c|d)$

Régression logistique vue de très loin

Vu de loin

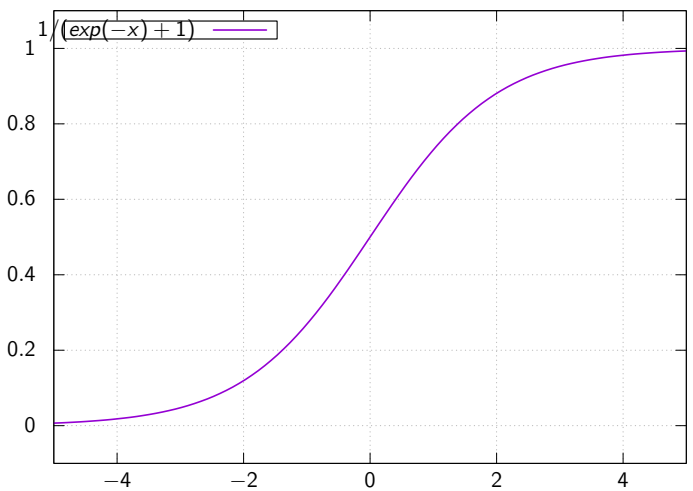
- Étant donné des données avec plusieurs *features* x_i
- Prédit une classe y
- Rien d'autre qu'une fonction linéaire sur les x_i
- Applique une sigmoïde pour garder le résultat du calcul entre 0 et 1



Régression logistique vue de très loin

Vu de loin

- Étant donné des données avec plusieurs *features* x_i
- Prédit une classe y
- Rien d'autre qu'une fonction linéaire sur les x_i
- Applique une sigmoïde pour garder le résultat du calcul entre 0 et 1
- *Le reste c'est du détail technique pour faire marcher le bousin*



Régressions logistiques

Autres aspects qui font marcher le truc

- La fonction de coût *cross-entropy*
  - son *log* marche bien avec l'*exp* de la sigmoïde
- La descente de gradient
  - Manière d'ajuster les poids (pas la seule)
- Variante stochastique de la descente de gradient
  - Plus rapide et facile à calculer (tient en mémoire)
- Variante mini-batch
  - Convergence plus lisse

Régressions logistiques pour le NLP

Appliquées sur le Bag-of-Words

- Les features sont les fréquences d'apprition des mots
- Potentiellement tronquées à 1
- Potentiellement tronquées à k
- Potentiellement *« écrasées »* avec un log

Questions ?

Questions ?

- Questions ?

TP

TP

- TP