Pastiche detection based on stopword rankings

Exposing impersonators of a Romanian writer

Liviu P. Dinu[1, 3]
ldinu@fmi.unibuc.ro

Vlad Niculae[1, 3]
vlad@vene.ro

Octavia-Maria Şulea[1, 2, 3]
mary.octavia@gmail.com

University of Bucharest

[1] Faculty of Mathematics and Computer Science [2] Faculty of Foreign Languages and Literatures [3] Center for Computational Linguistics

About pastiche detection

Pastiche: exercise in copying another author's style.
Sometimes overt, sometimes **deceptive**: the impersonator is passing off his own work as another's.

Approach as a **computational stylometry** problem: look at the stopword fingerprint (Somers & Tweedie, 2003).

In Romanian literature: *the successors of Mateiu Caragiale*. Includes deceptive pastiches, overt pastiches and stylistically similar works.

Comparing rankings. Rank distance

How similar are two rankings: difficult question.

Spearman's footrule for permutations is too specific:
In a ranking, two entries can have the same rank.
Two rankings can have different entries.

Extension: **Rank distance** (Dinu)
Found uses in computational biology and NLP

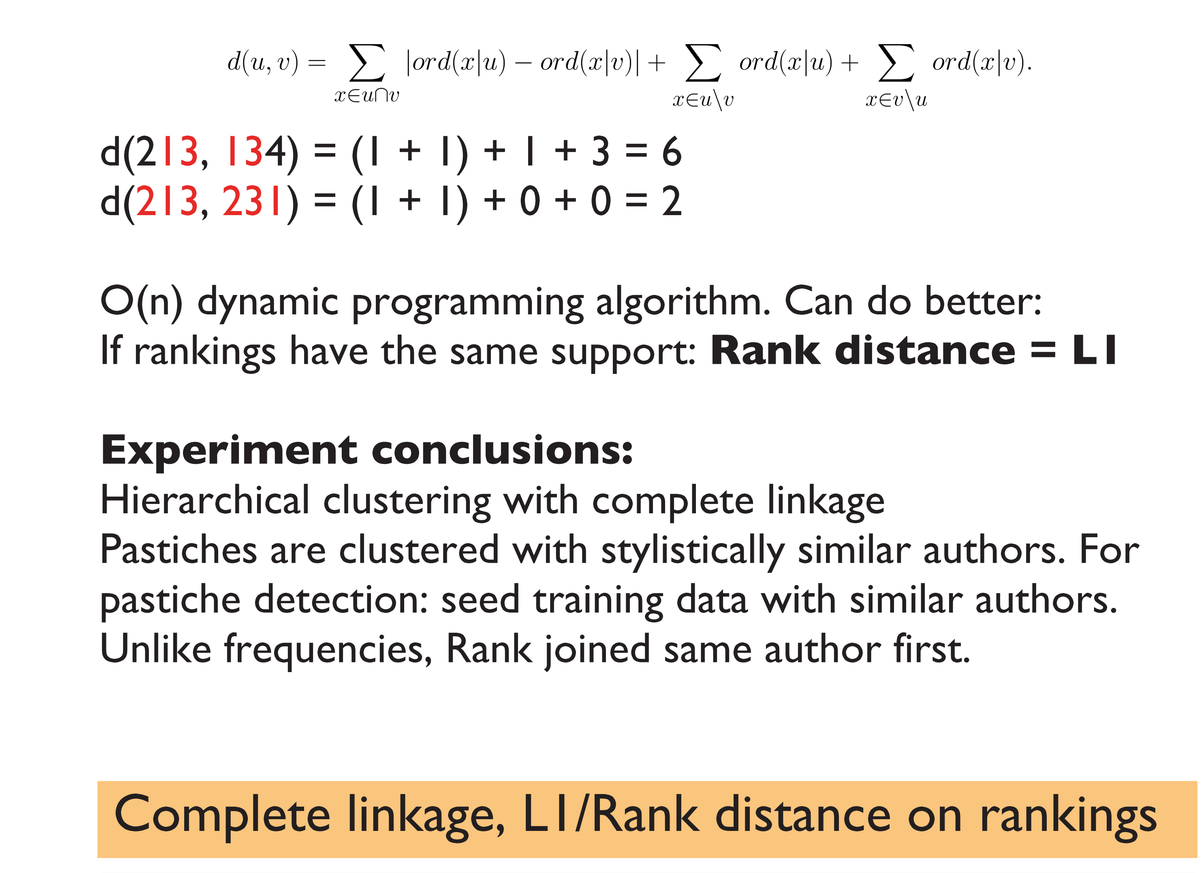
$$d(u,v) = \sum_{x \in u \cap v} |ord(x|u) - ord(x|v)| + \sum_{x \in u \setminus v} ord(x|u) + \sum_{x \in v \setminus u} ord(x|v).$$

d(213, 134) = (1 + 1) + 1 + 3 = 6
d(213, 231) = (1 + 1) + 0 + 0 = 2

O(n) dynamic programming algorithm. Can do better:
If rankings have the same support: **Rank distance = L1**

Experiment conclusions:
Hierarchical clustering with complete linkage
Pastiches are clustered with stylistically similar authors. For pastiche detection: seed training data with similar authors.
Unlike frequencies, Rank joined same author first.

Complete linkage, L1/Rank distance on rankings



Stopword rankings as features

Standard approach is based on stopword frequencies. Do we need the exact frequency?

Somers & Tweedie found L2 clustering of frequencies to be good but imperfect. We observed the same issue.

Reason: exact frequencies are too sensitive. We care only about the **order in which the stopwords are more likely to occur**. This acts as **smoothing**.

frequencies:

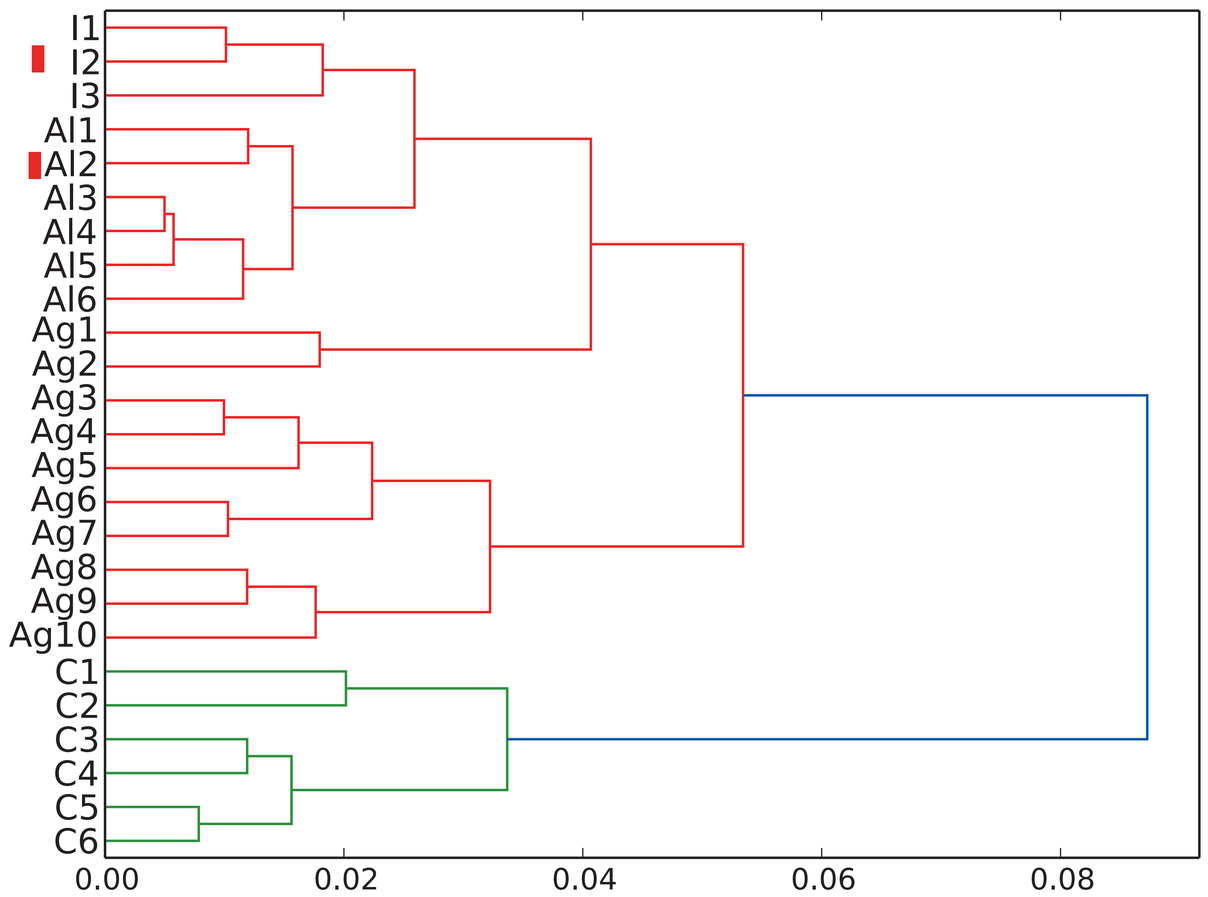
'a':	2342
'about':	618
'above':	372
...	
'that':	1013
'the':	3360
...	

sort, dismiss exact values

ranking:

1.	'the'
2.	'a'
3.	'that'
...	
17.	'about'
...	
23.	'above'
...	

Complete linkage, L2 distance on frequencies



Literary works used in this experiment

Highlighted works are admitted pastiches

Ion Iovan:
I1: Epistolar
I2: Ultimele însemnări ale lui Mateiu Caragiale
I3: Indexul fiinţelor, lucrurilor şi întâmplărilor

Radu Albala:
Al1: Propyläen Kunstgeschichte
Al2: În deal, pe Militari
Al3: La Paleologu
Al4: Sclava iubirii
Al5: Femeia de la miezul nopţii
Al6: Nişte cireşe

Ştefan Agopian:
Ag1: Republica pe eşafod (drama)
Ag2: Drumul (drama)
Ag3: Manualul întâmplărilor (sort of a novel)
Ag4: Însemnări din Sodoma (sort of a novel)
Ag5: Ziua Mâniei (novel)
Ag6: Tache de catifea (novel)
Ag7: Manualul întâmplărilor (drama)
Ag8: Tobit
Ag9: Sara
Ag10: Frică

Eugen Bălan:
Bln: Sub pecetea tainei

Mateiu Caragiale:
C1: Studii heraldice
C2: Pajere (1936)
C3: Jurnal (1927-1935)
C4: Remember (1924)
C5: Sub pecetea tainei (1930, 1933)
C6: Craii de Curtea-Veche (1926)