NLP 2025 - Exercise
Information Retrieval Basics

---

1. Inverted Index

สมมติว่ามีเอกสารสี่ชิ้นซึ่งมีคำดังต่อไปนี้

Doc 1: new home in top forecasts

Doc 2: home sales rise in july

Doc 3: increase in home sales in july

Doc 4: rise in new home sales

จงวาด Inverted Index สำหรับ boolean retrieval ที่สามารถตอบสนอง phrase query ได้โดยไม่ต้องใช้ bigram index (แปลว่าต้องเก็บตำแหน่งที่แต่ละคำเกิดขึ้น เอาไว้ใน postings ด้วย) จะทำด้วยมือหรือเขียนโค้ดก็ได้ โดยเริ่มจาก

```
doc1 = 'new home in top forecasts'.split(' ')
doc2 = 'home sales rise in july'.split(' ')
doc3 = 'increase in home sales in july'.split(' ')
doc4 = 'rise in new home sales'.split(' ')
docs = [doc1, doc2, doc3, doc4]
```

Inverted Index ที่วาดควรมีลักษณะดังนี้

Term	Postings
new	(1, [0]), (4,[2])
...	
...	

2. Ranked Retrieval

สมมติว่าเราได้ term-doc matrix ดังตารางข้างล่าง

term	Doc 1	Doc 2	Doc 3
Linus	10	0	1
Snoopy	1	4	0
pumpkin	4	100	10

ถ้า query = "Linus pumpkin" และใช้ TF-IDF เป็นเกณฑ์คะแนนความเกี่ยวข้อง (Relevance score) ดังสูตรข้างล่างโดยที่ใช้ log ฐาน 10 และ N คือจำนวน document ทั้งหมด

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

ผลการค้นหาจะออกมาเป็นอย่างอะไร ให้จัดอันดับของแต่ละ document ตามคะแนนความเกี่ยวข้อง และแสดงวิธีการคำนวณ TF-IDF ด้วย
(คำใบ้: log 1 = 0 หา IDF ของแต่ละ term ก่อนแล้วจะง่ายมาก)