Laurea Magistrale in
Ingegneria Matematica
a.a. 2019/2020

POLITECNICO
DI TORINO

Progetto di Business Intelligence per i Big Data

Luca Bajardi
Lidia Fantauzzo

Obiettivi

- Clusterizzazione delle news mediante analisi testuale
- Caratterizzazione dei cluster in base all'influenza delle news sull'andamento finanziario degli stock di appartenenza

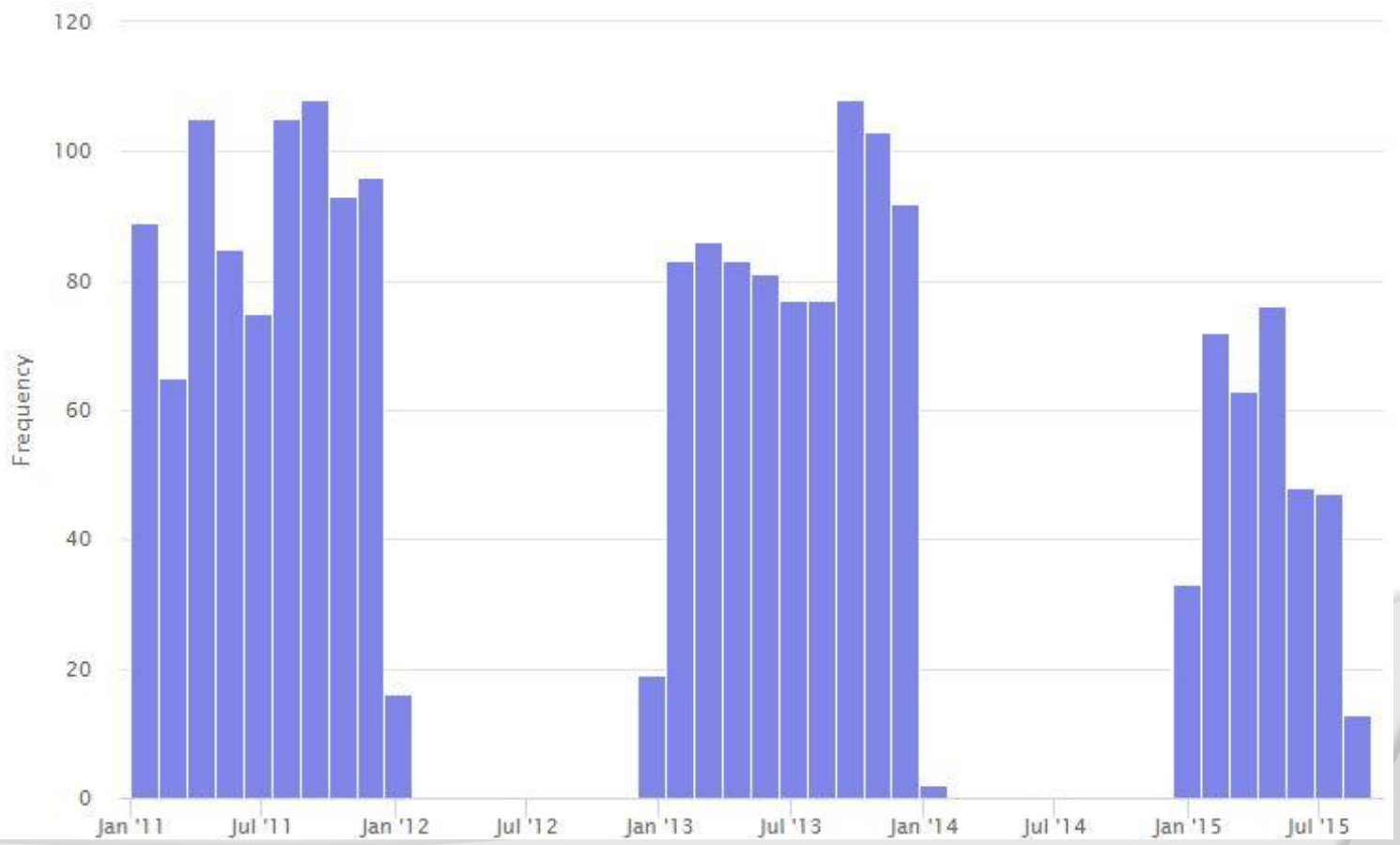
Esplorazione dei dati

- **Schema del dataset:**
  - Date
  - Body
  - Stock
  - Positive
  - Negative

Row No.	date	body	stock	positive	negative
1	May 29, 2015	(The followin...	T	8	14
2	Oct 21, 2013	WASHINGTO...	AMZN	2	10
3	Jul 16, 2013	NEW YORK (...	GOOGL	3	20
4	Apr 12, 2013	(Members of ...	TWTR	1	23
5	Mar 28, 2015	March 27 (Re...	GS	0	0
6	Nov 4, 2013	Nov 4 (Reuter...	TWTR	2	2
7	Mar 16, 2011	TOKYO, Mar ...	TWTR	1	3

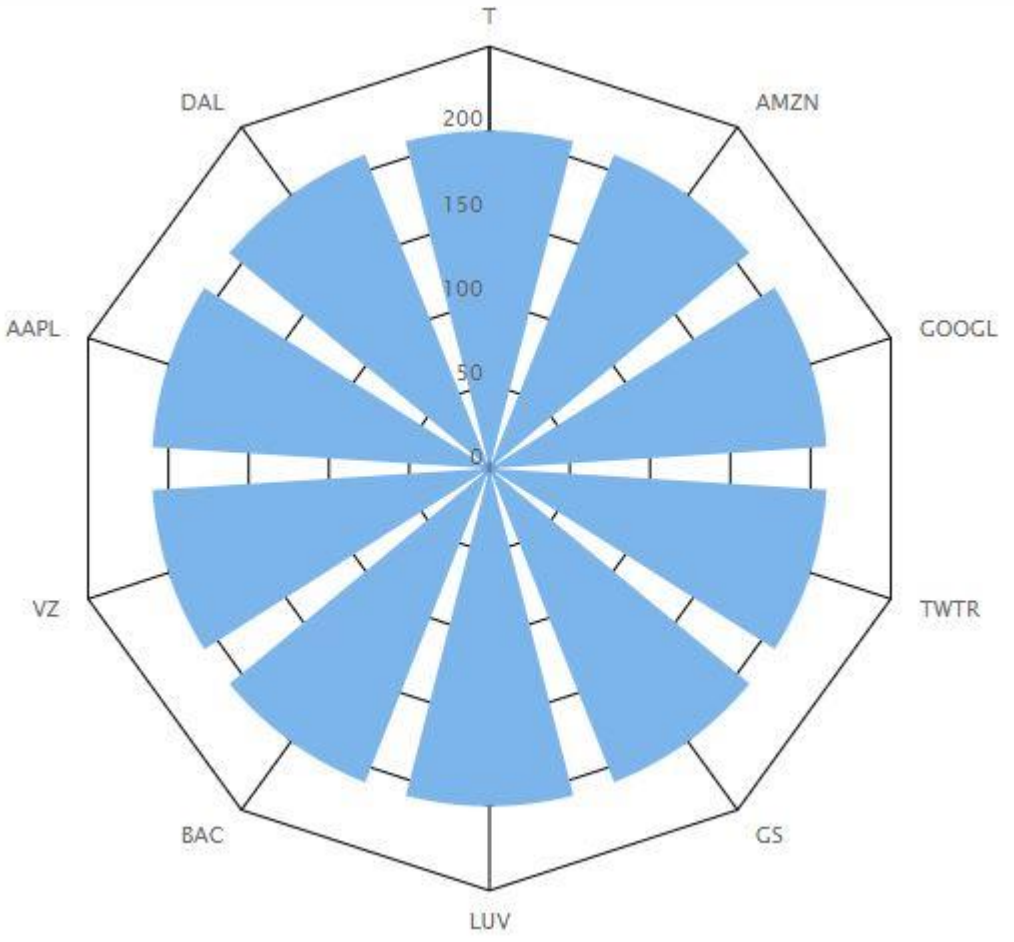
Esplorazione dei dati

- **Distribuzione delle date:**



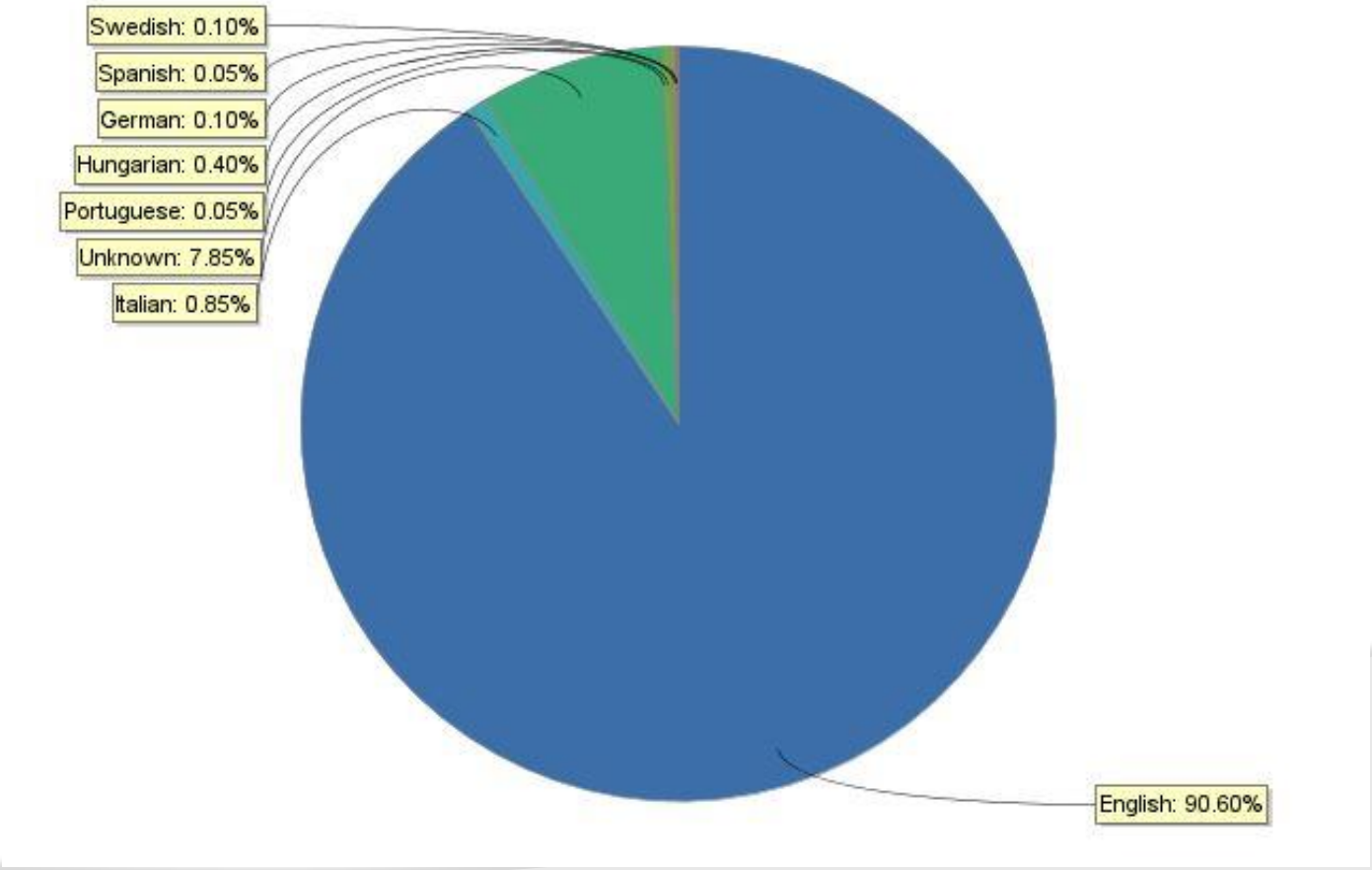
Esplorazione dei dati

- **Distribuzione degli stock:**



Esplorazione dei dati

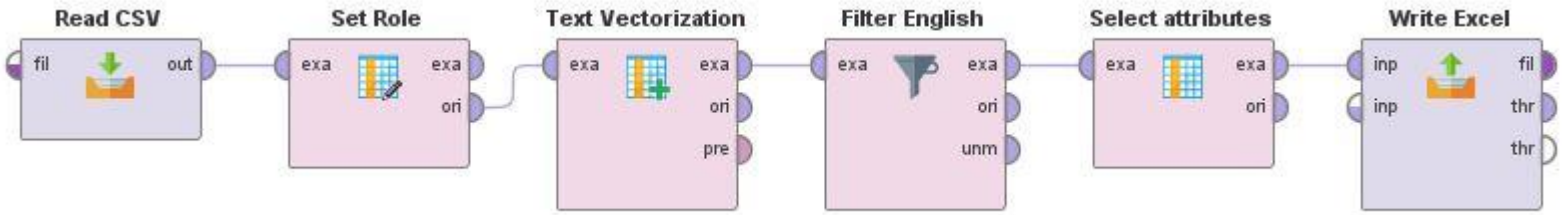
- **Distribuzione delle lingue delle news:**



Pre-processing

- Eliminazione delle news non in lingua inglese

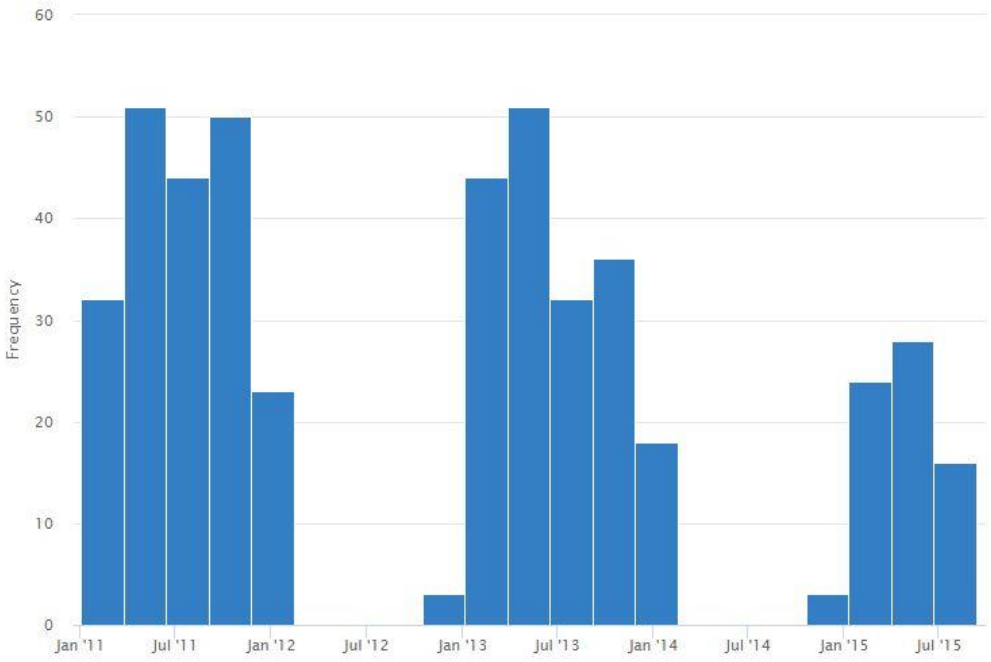
Processo di RapidMiner



- 1812 istanze filtrate
- Scrittura di un nuovo file per ottimizzare il processo

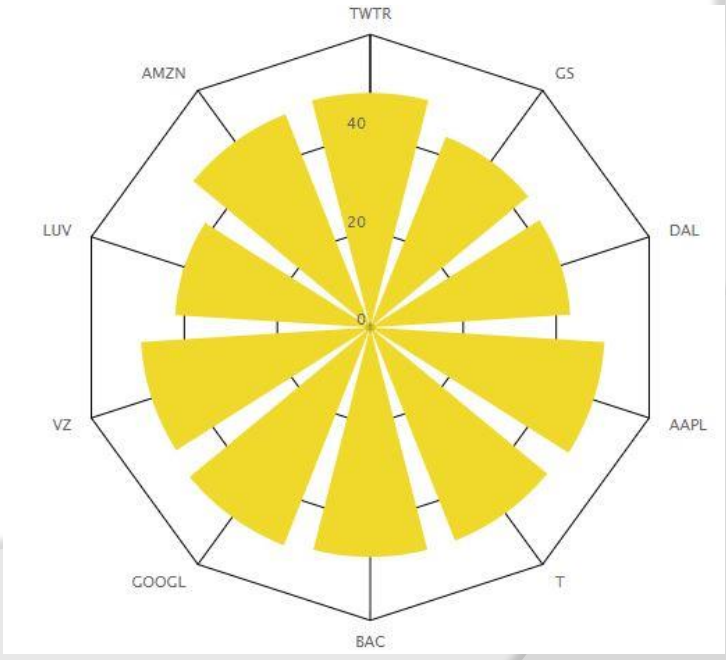
Campionamento

- Stratificato rispetto alla variabile stock



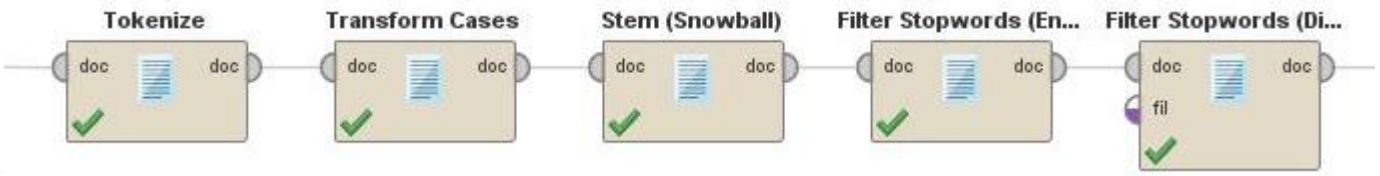
Campione del 25%:
455 istanze

Valori tra 40-48

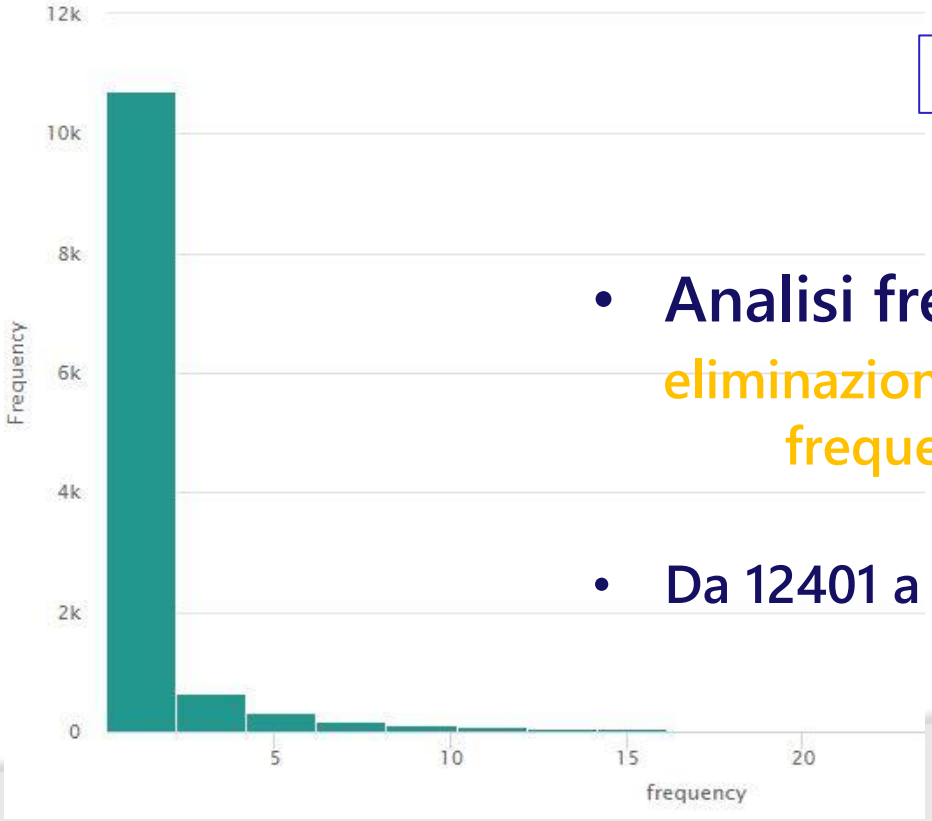


Pre-processing

- Filtraggio delle parole nei documenti



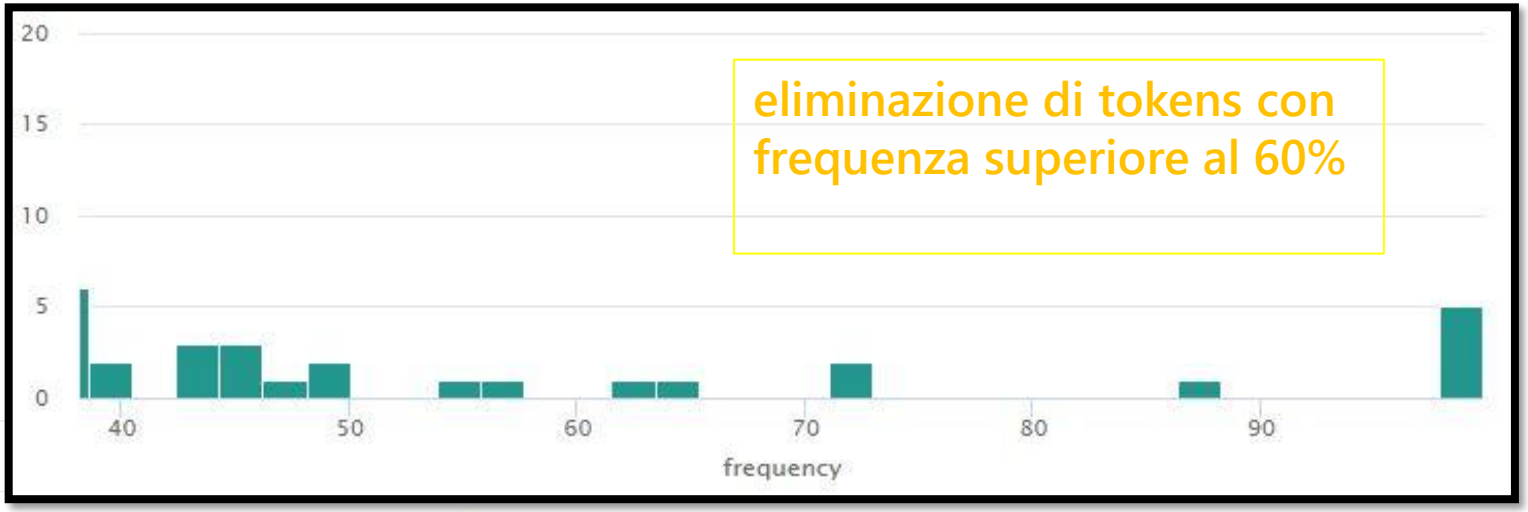
Doppio filtro stopwords



- Analisi frequenza tokens - step1:
  eliminazione di tokens con frequenza inferiore a 4,2%
- Da 12401 a 1074 attributi

Pre-processing

- Analisi frequenza tokens - step2:



Zoom

- Da 1074 a 1064 attributi

Clustering dei documenti

- **Struttura del processo:**
  - **Scelta del numero di dimensioni SVD:**
    - **350 (90%)**
    - **251 (70%)**
    - **150 (55%)**
  - **Scelta del metodo di clustering:**
    - **DBSCAN**
    - **K-Means**
  - **Scelta di due misure da confrontare:**
    - **Distanza euclidea**
    - **Similarità del coseno**

Clustering: DBSCAN misura del coseno

Total Cohesion

$$TC = \sum_i \sum_{x \epsilon C_i} \frac{x \cdot m_i}{\|x\| \cdot \|m_i\|}$$

Id	Dim	Eps	Min points	TC
1	350	1,49	8	2,74
2	251	1,46 1,49	15 15	14,46
3	251	1,46 1,495	16 19	12,15
4	150	1,41 1,45	20 20	13,94
5	150	1,42 1,47	20 16	23,79

Scelta modello 5:

Prima clusterizzazione:

```
Cluster 0: 164 items
Cluster 1: 191 items
Cluster 2: 21 items
Cluster 3: 33 items
Cluster 4: 26 items
Cluster 5: 20 items
Total number of items: 455
```

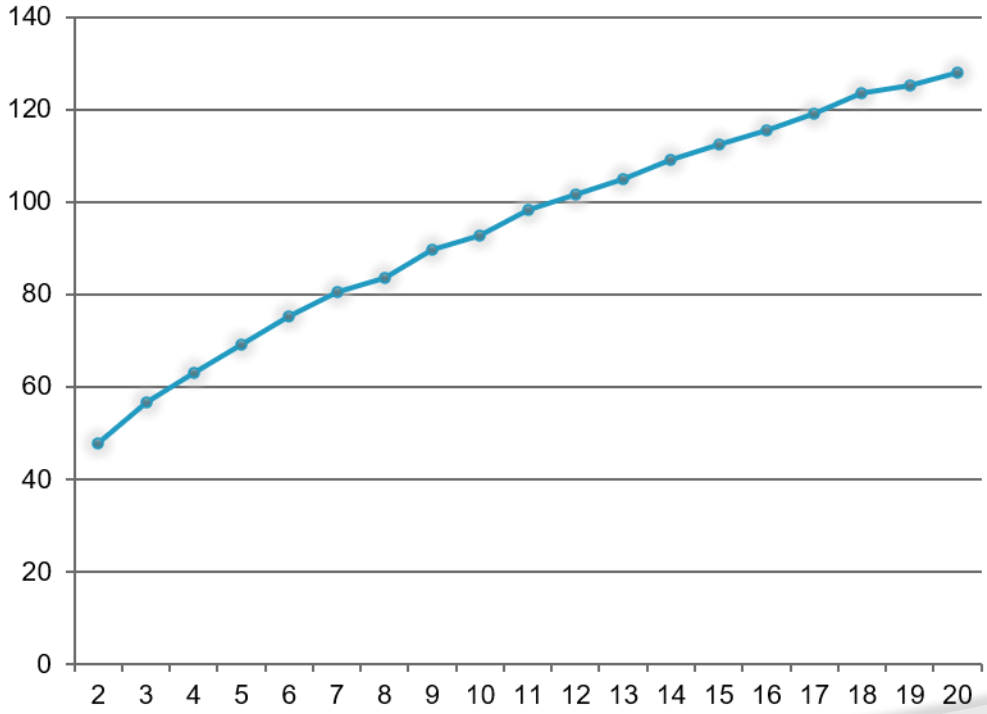
Seconda clusterizzazione:

```
Cluster 0: 44 items
Cluster 1: 65 items
Cluster 2: 11 items
Cluster 3: 12 items
Cluster 4: 16 items
Cluster 5: 16 items
Total number of items: 164
```

Clustering: K-Means misura del coseno

- **Eliminazione di 44 outliers e applicazione del metodo:**

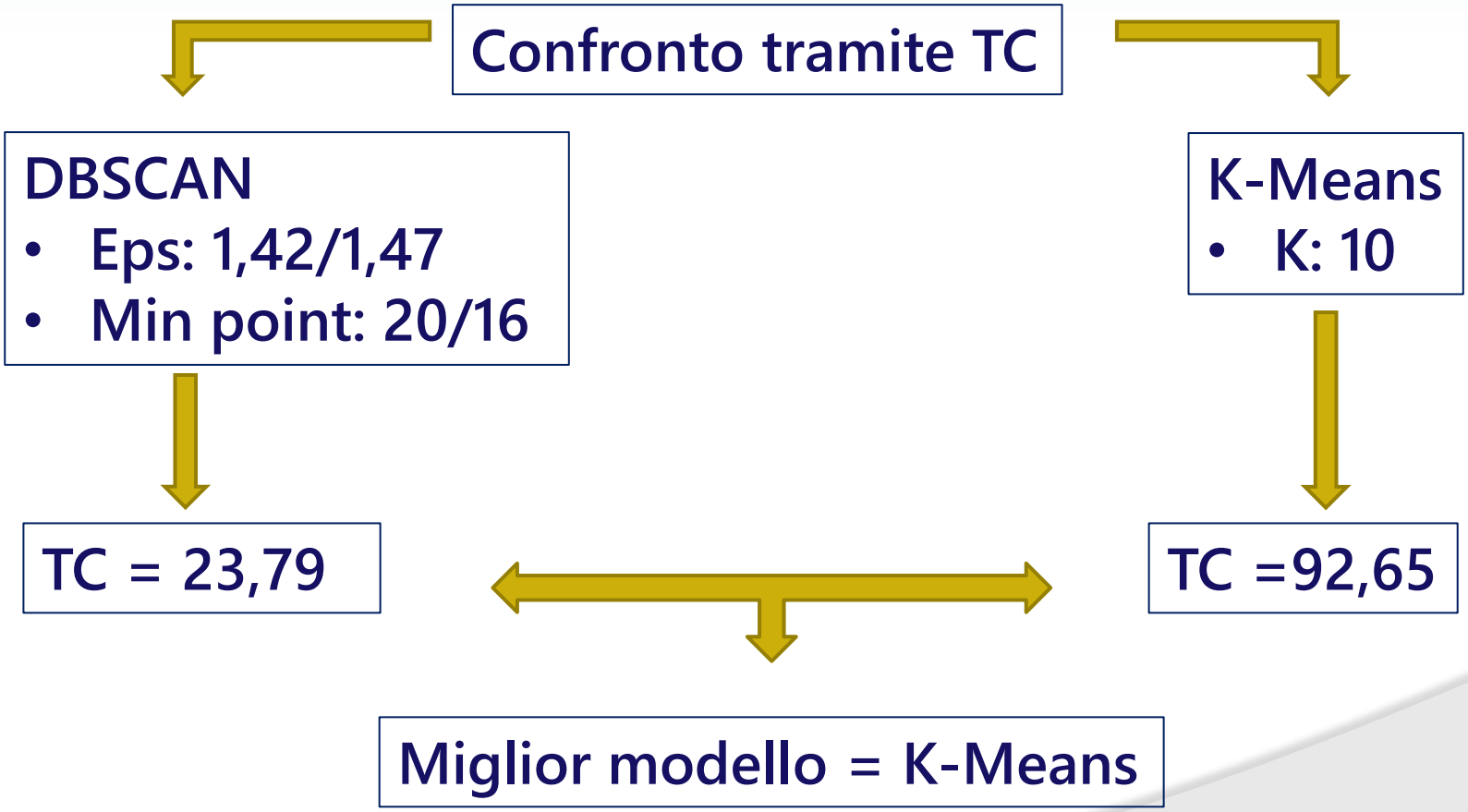
Grafico TC al variare di k



k=10 TC=92,65

```
Cluster 0: 38 items
Cluster 1: 49 items
Cluster 2: 38 items
Cluster 3: 42 items
Cluster 4: 39 items
Cluster 5: 44 items
Cluster 6: 37 items
Cluster 7: 34 items
Cluster 8: 51 items
Cluster 9: 39 items
Total number of items: 411
```

Clustering con misura del coseno: miglior modello



Clustering: DBSCAN misura euclidea

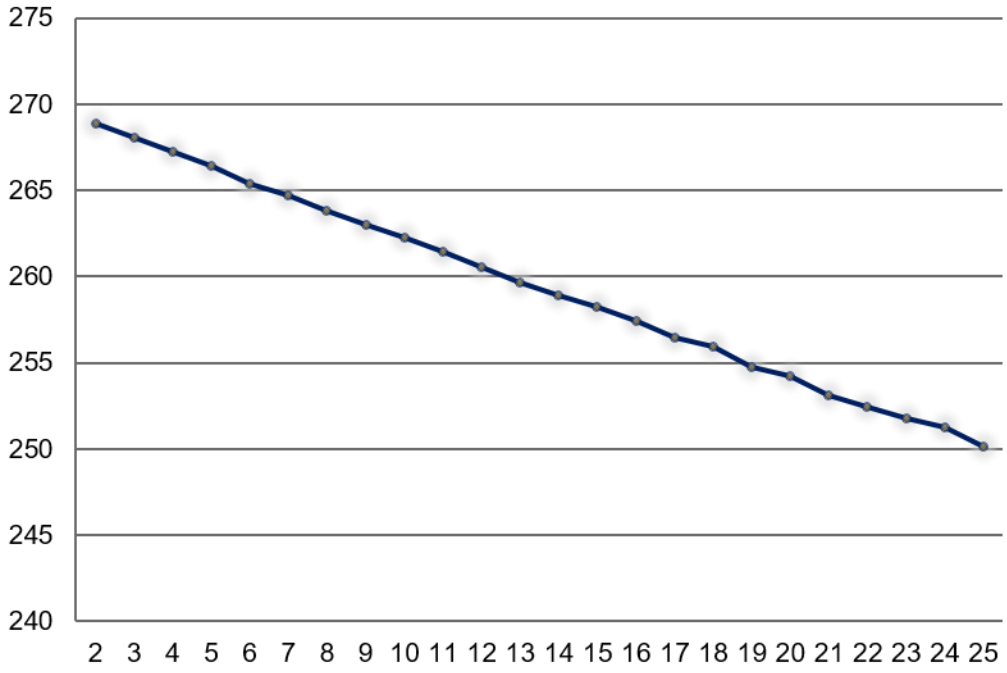
- Tre dimensioni:
  - non direttamente confrontabili tramite SSE
  - migliori parametri: grafico k-distances
  - Modello scelto: dimensione 350, per la maggiore varianza spiegata (90%)

Dimensione	Eps	Min points	SSE	Clusters
350	1,1	7	269.79	0: 88 1:367
251	0,97	7	215,82	0:46 1:409
150	0,7	6	120,46	0:66 1:389

Clustering: K-Means misura euclidea

- Eliminazione di 88 outliers e applicazione del metodo:

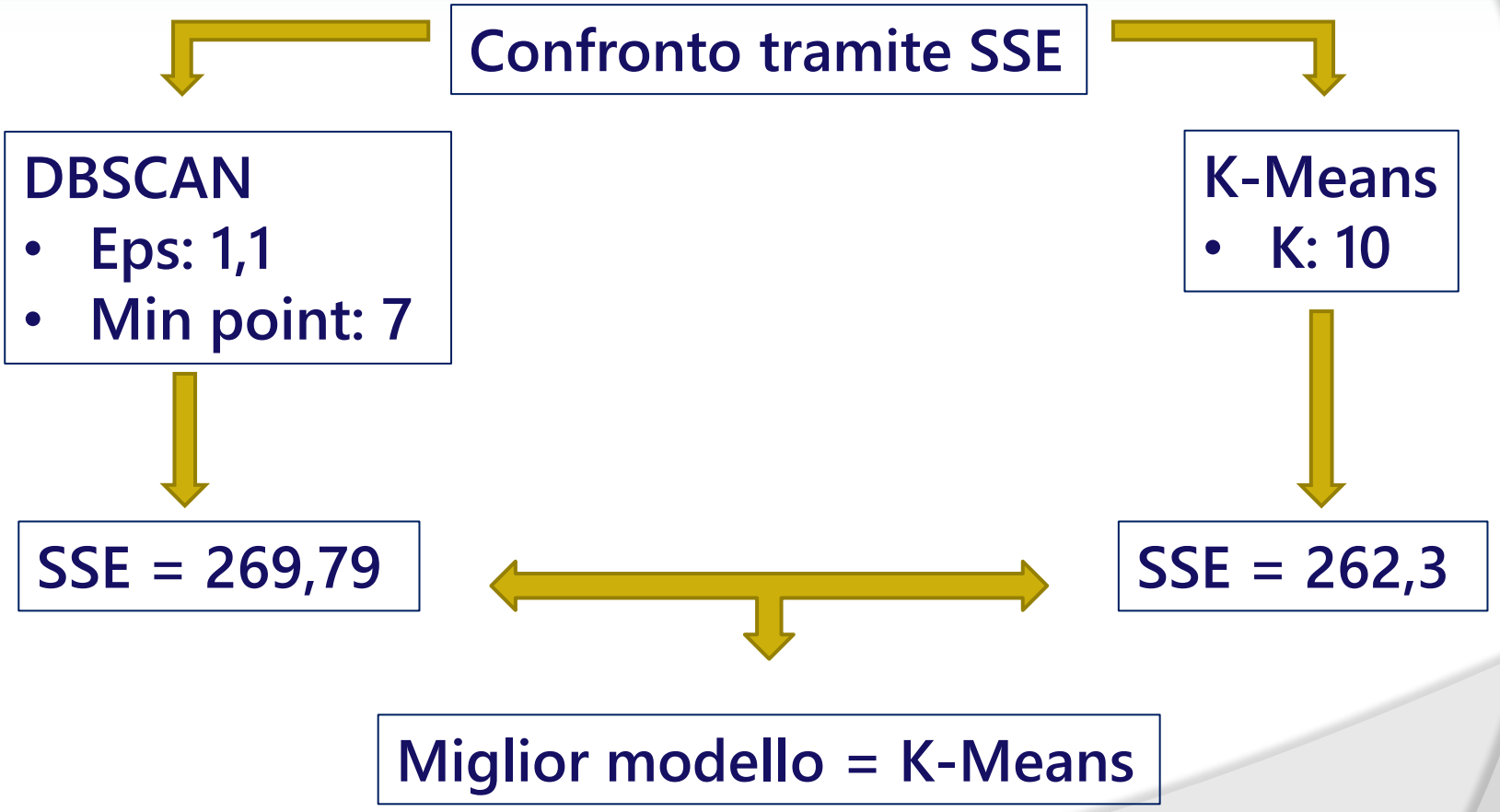
Grafico SSE al variare di k



k=10 SSE=262,30

```
Cluster 0: 11 items
Cluster 1: 2 items
Cluster 2: 5 items
Cluster 3: 9 items
Cluster 4: 331 items
Cluster 5: 2 items
Cluster 6: 1 items
Cluster 7: 3 items
Cluster 8: 2 items
Cluster 9: 1 items
Total number of items: 367
```

Clustering con misura euclidea: miglior modello



Confronto tra i due migliori

K-Means
- **K: 10**
- **Misura euclidea**

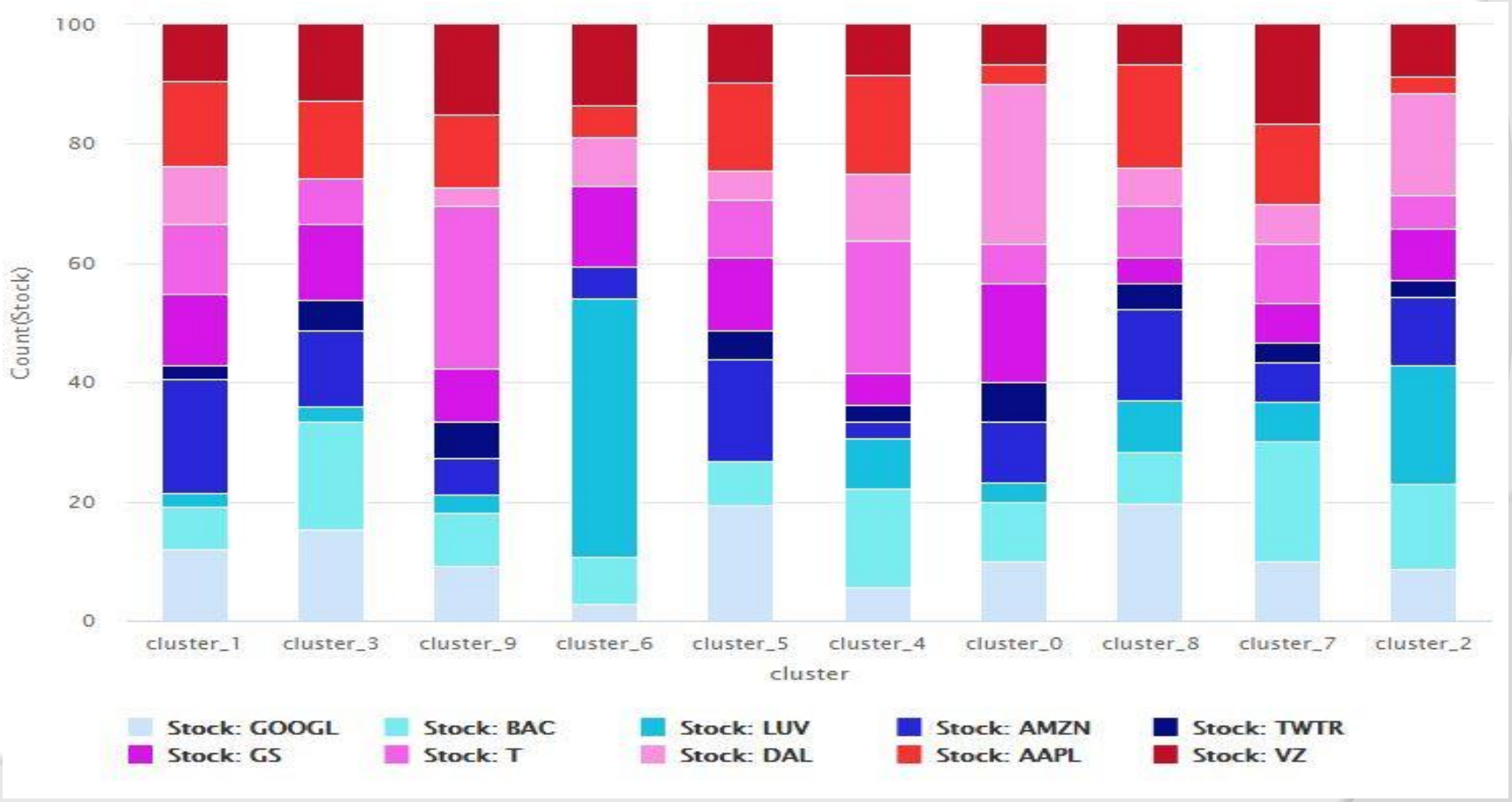
K-Means
- **K: 10**
- **Misura del coseno**

Maggiore interpretabilità e omogeneità dei cluster

K-Means con misura del coseno

Caratterizzazione clusters

Distribuzione degli stock: **omogenea**



Integrazione dataset

https://markets.financialcontent.com/stocks/quote/

Date	Open	High	Low	Close	Volume	Change (%)
Jun 02, 2020	320.74	323.44	318.93	323.34	21,865,005	+1.49(+0.46%)
Jun 01, 2020	317.75	322.35	317.21	321.85	20,228,555	+3.91(+1.23%)
May 29, 2020	319.25	321.15	316.47	317.94	38,399,500	-0.31(-0.10%)
May 28, 2020	316.77	323.44	315.63	318.25	33,418,268	+0.14(+0.04%)

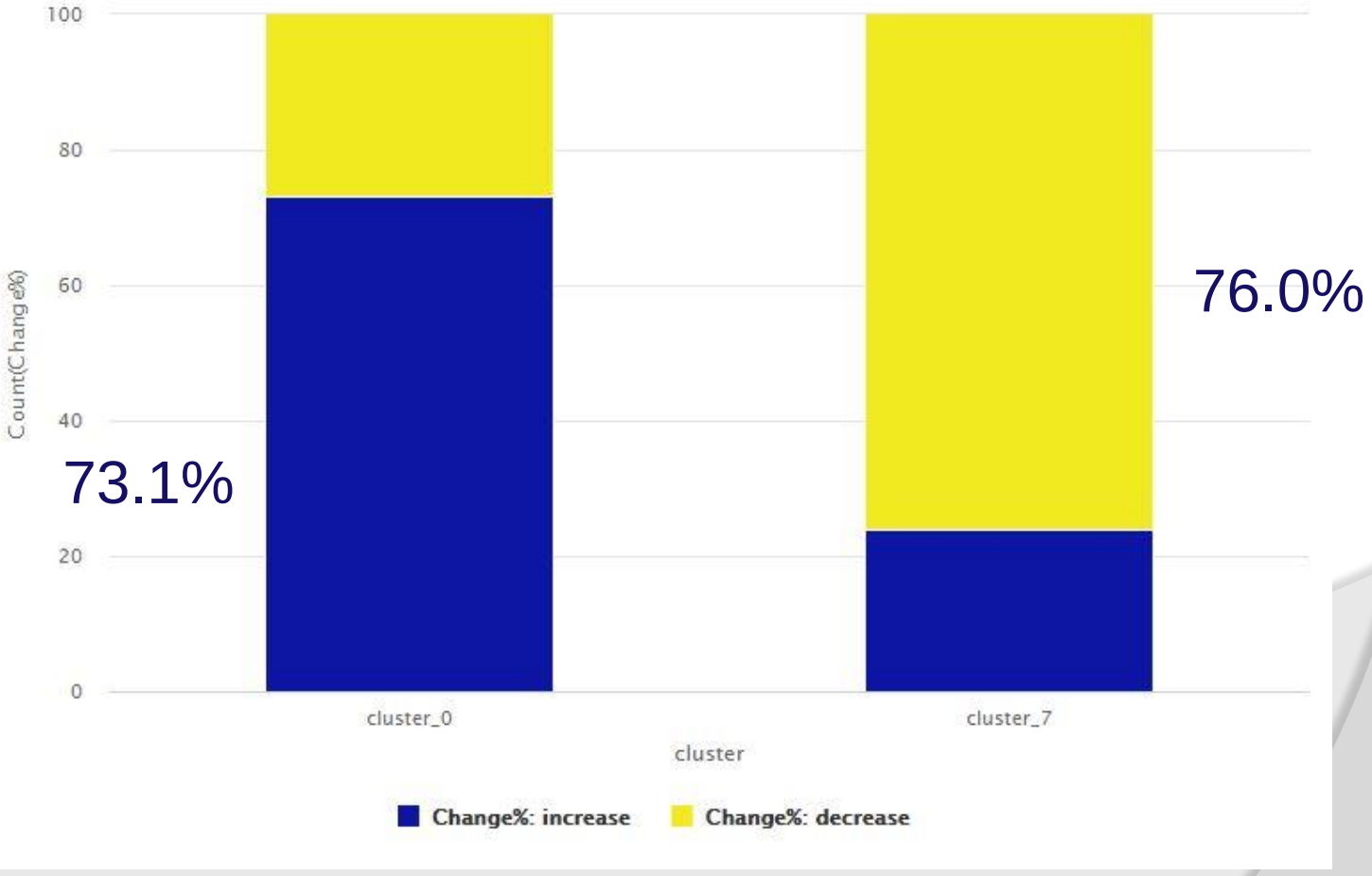
- Ridatazione delle news pubblicate nel weekend
- Union di tutti i dataset per ogni stock
- Discretizzazione Change%:
  - decrease: (-inf,0.0]
  - increase: (0.0,+inf)

Caratterizzazione clusters

Cluster	Carattere
0	increase
1	increase
2	decrease
3	increase
4	increase
5	increase
6	decrease
7	decrease
8	increase
9	decrease

Caratterizzazione clusters

Analisi Change%:



Caratterizzazione clusters: Regole di associazione

Cluster_0 carattere: **‘increase’**

Regola	Supporto	Confidenza	Lift	Conviction
accord → expect	0,43	0,93	1,64	6,06
trade → bank, market, accord	0,3	0,75	2,25	2,67
bank, accord → billion	0,3	0,75	1,41	1,87
bank, market, growth →money	0,2	0,75	3,21	3,07

Caratterizzazione clusters: Regole di associazione

Cluster_7 carattere: **‘decrease’**

Regola	Supporto	Confidenza	Lift	Conviction
commiss → dial	0,27	0,6	1,65	1,59
lawsuit → court, judg	0,21	0,7	3,3	2,62
financi → declin	0,21	0,64	2,63	2,08

Grazie per l’attenzione