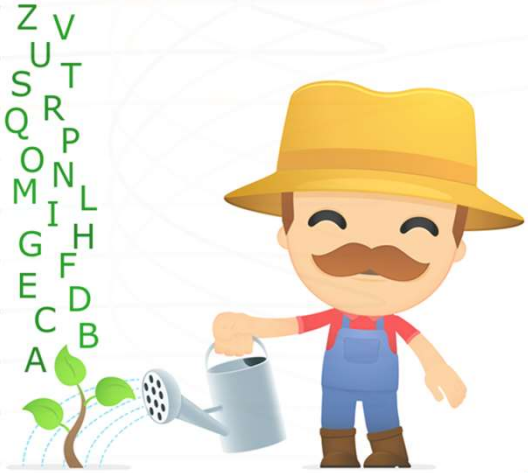
crea
Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria

Centro di politiche e Bioeconomia
Ufficio di Statistica

Introduzione all'*analisi automatica dei testi* e potenziali applicazioni in Agricoltura



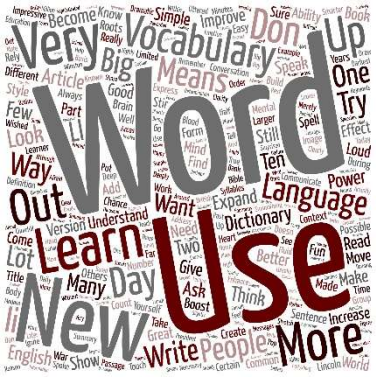
Marco Vassallo

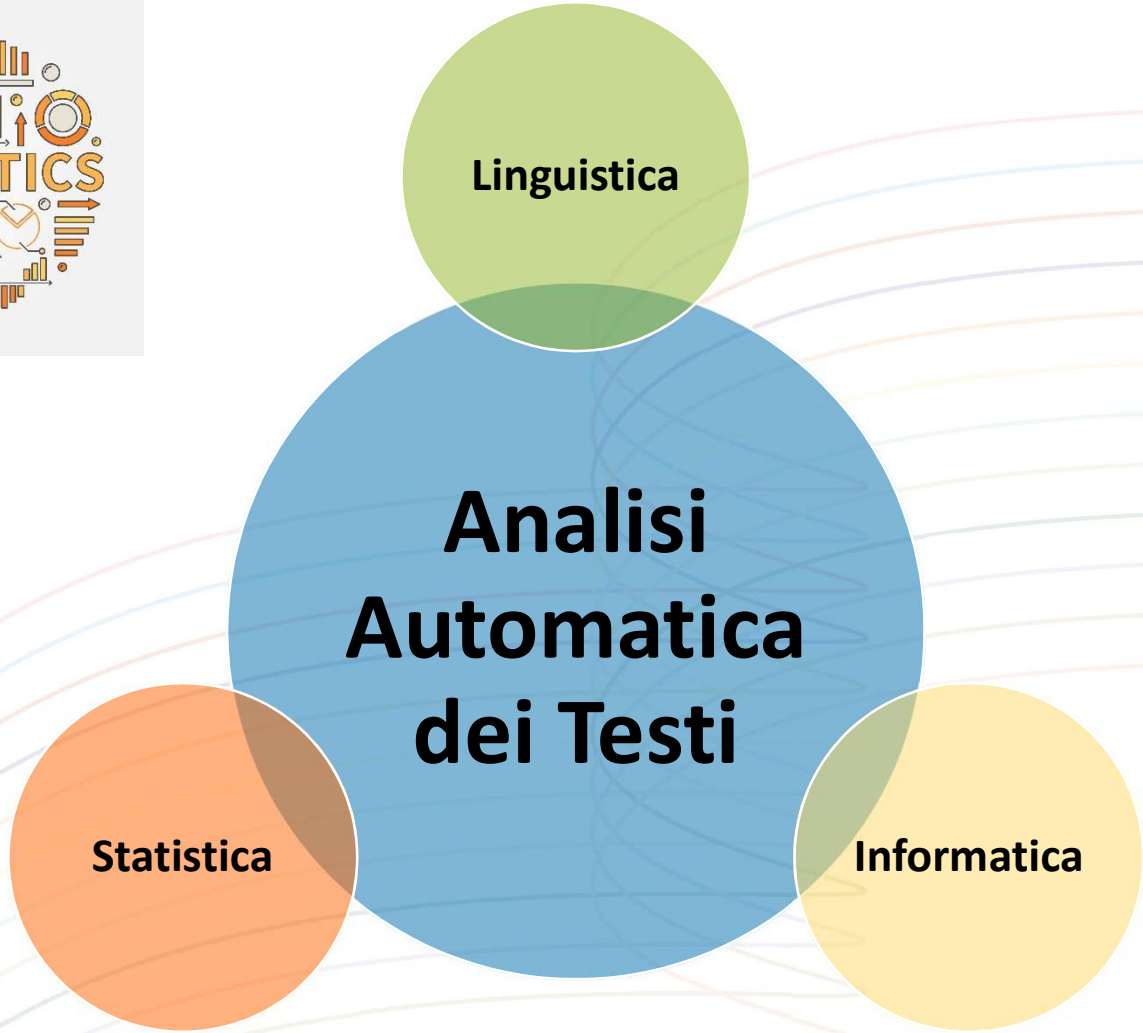
Roma, 19 febbraio 2019

Sommario

- Cos'è l'analisi automatica dei testi?
- Cosa vuol dire analizzare automaticamente un testo?
- Potenziali applicazioni in Agricoltura
- Prospettive future

Cos'è l'Analisi Automatica dei Testi?





Cos'è l'Analisi Automatica dei Testi?

Già nel 1995 un rapporto della Forrester Research* stimava che l'80% delle informazioni in un'azienda era costituita da testi ed il 20% da dati numerici…

**Forrester Research (1995). Coping with Complex Data, The Forrester Report, April. In: Giuliano e La Rocca (2008), L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso. Milano, LED, pag. 156.*

Cos'è l'Analisi Automatica dei Testi?

E nel 2020

Bernard Marr (2015) dichiara su Forbes che nel 2020 ogni essere umano produrrà 1.7 megabyte di dati ogni secondo…

Cos'è l'Analisi Automatica dei Testi?

Javad Zarif ✓
@JZarif

Segui

Iran took & concluded proportionate measures in self-defense under Article 51 of UN Charter targeting base from which cowardly armed attack against our citizens & senior officials were launched.

We do not seek escalation or war, but will defend ourselves against any aggression.

18:32 · 7 gen 2020

20.060 Retweet **55.767** Mi piace

9899 20060 55767

Donald J. Trump ✓
@realDonaldTrump

Segui

All is well! Missiles launched from Iran at two military bases located in Iraq. Assessment of casualties & damages taking place now. So far, so good! We have the most powerful and well equipped military anywhere in the world, by far! I will be making a statement tomorrow morning.

18:45 · 7 gen 2020

159.500 Retweet **767.565** Mi piace

151692 159500 767565

Cosa vuol dire analizzare automaticamente un testo?

Testo non strutturato

Analisi Lessicale Automatica

Testo strutturato

Text Analysis

Opinion Mining

Text Mining

Potenziali applicazioni in Agricoltura

Focus Group – Text Mining

IRaMuTeQ

R Interface for multidimensional analysis of texts and questionnaires

R & python

http://www.iramuteq.org/

CREAgritrend – Opinion Mining

https://www.r-project.org/

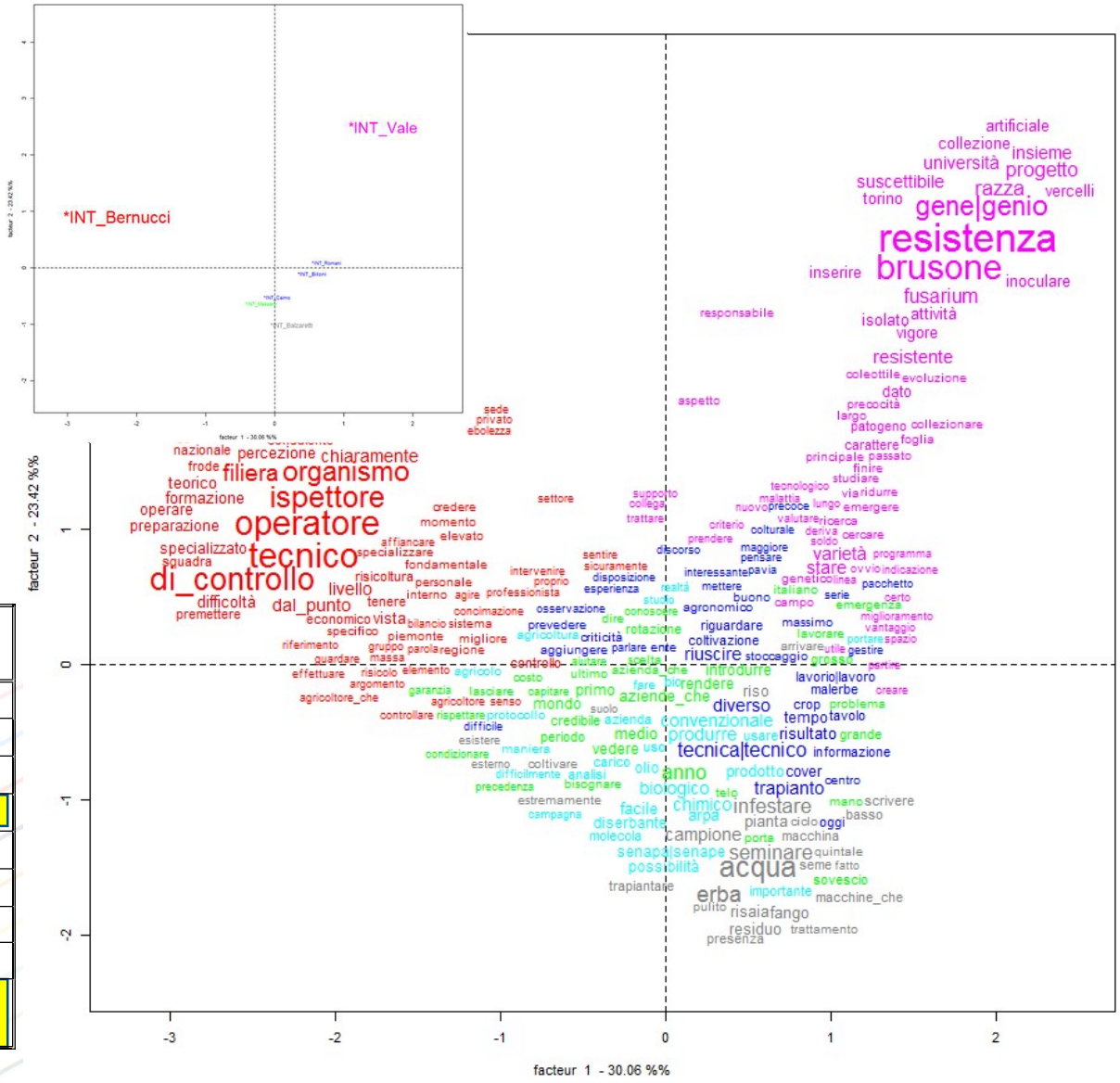
Partecipazione a CliC-it 2019

Wordcloud preso da: Monda M., Gabrieli G., & Vassallo M. (2019). Sentiment in Agricoltura: Il termometro dell'Agricoltura. In CREAgritrend: I numeri dell'Agroalimentare Italiano. Bollettino trimestrale edito dal CREA, Centro Politiche e Bio-economia, n°2, I Semestre 2019. ISSN: 2612-6419.

Potenziali applicazioni: Focus Group – analisi trascrizioni

Riso Biologico

Statistiche descrittive lessicometriche	*Focus Riso*
Numero di testi (intervistati)	7
Numero di occorrenze (N)	13174
Numero di forme grafiche (V)	1937
Type/token ratio = (V/N)*100	14.7%
Frequenza media = N/V	6.80
G di Guiraud	16.87
Coefficiente α	1.25
Numero di hapax	934
Hapax % *sulle occorrenze (N)* *sulle forme (V)*	7.1% 48.2%

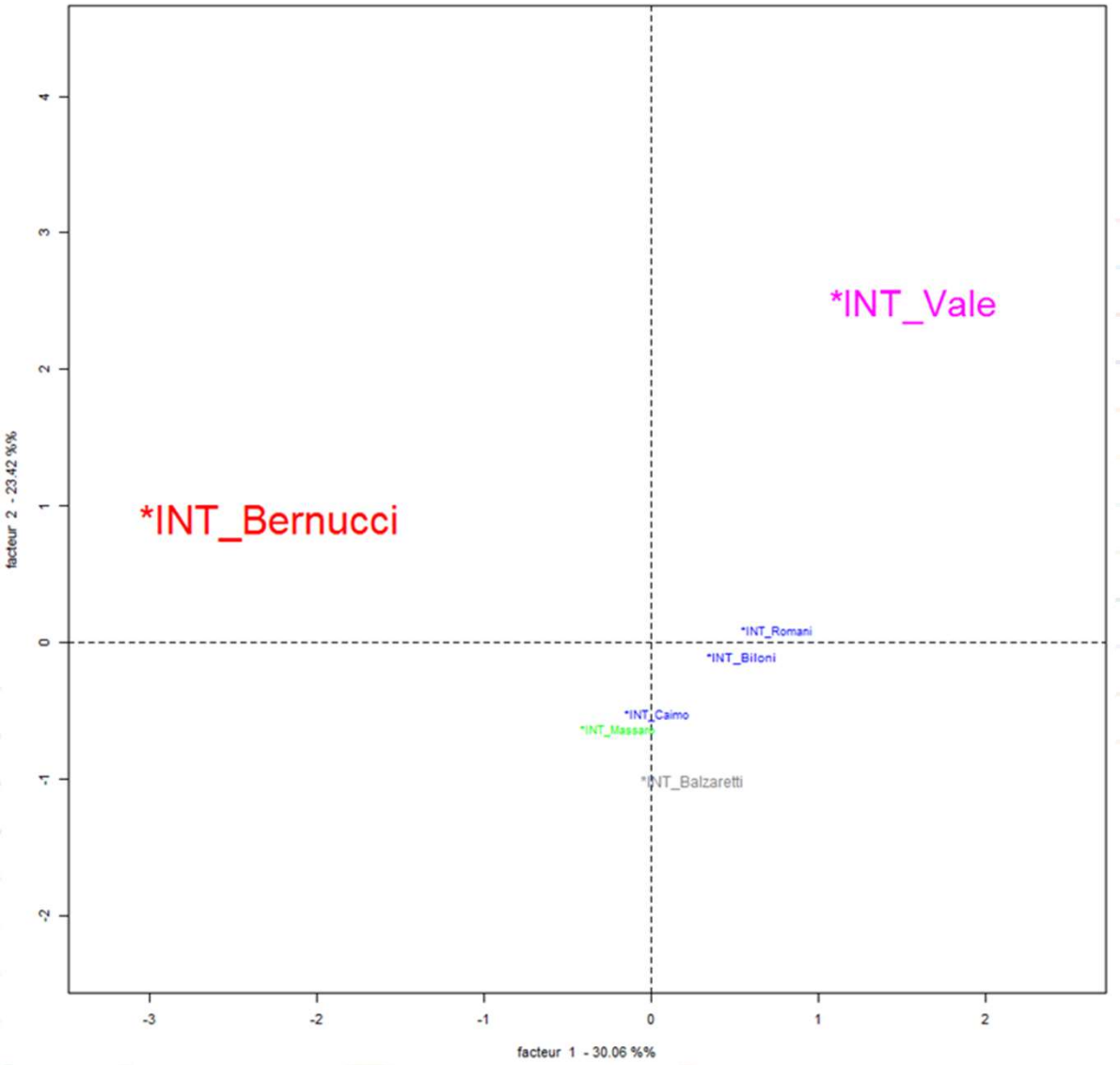


Guccione GD, Vaccaro A., Borsotto P., Vassallo M., Iacono R., Borri I. (2019). Risobiosystems: Documento di analisi del sistema di controllo e certificazione - Report WP3 Progetto di ricerca, sviluppo e trasferimento a sostegno della risicoltura biologica

Potenziali applicazioni: Focus Group – analisi trascrizioni

Riso Biologico

Statistiche descrittive lessicometriche	*Focus Riso*
Numero di testi (intervistati)	7
Numero di occorrenze (N)	13174
Numero di forme grafiche (V)	1937
Type/token ratio = (V/N)*100	14.7%
Frequenza media = N/V	6.80
G di Guiraud	16.87
Coefficiente α	1.25
Numero di hapax	934
Hapax % *sulle occorrenze (N)* *sulle forme (V)*	7.1% 48.2%

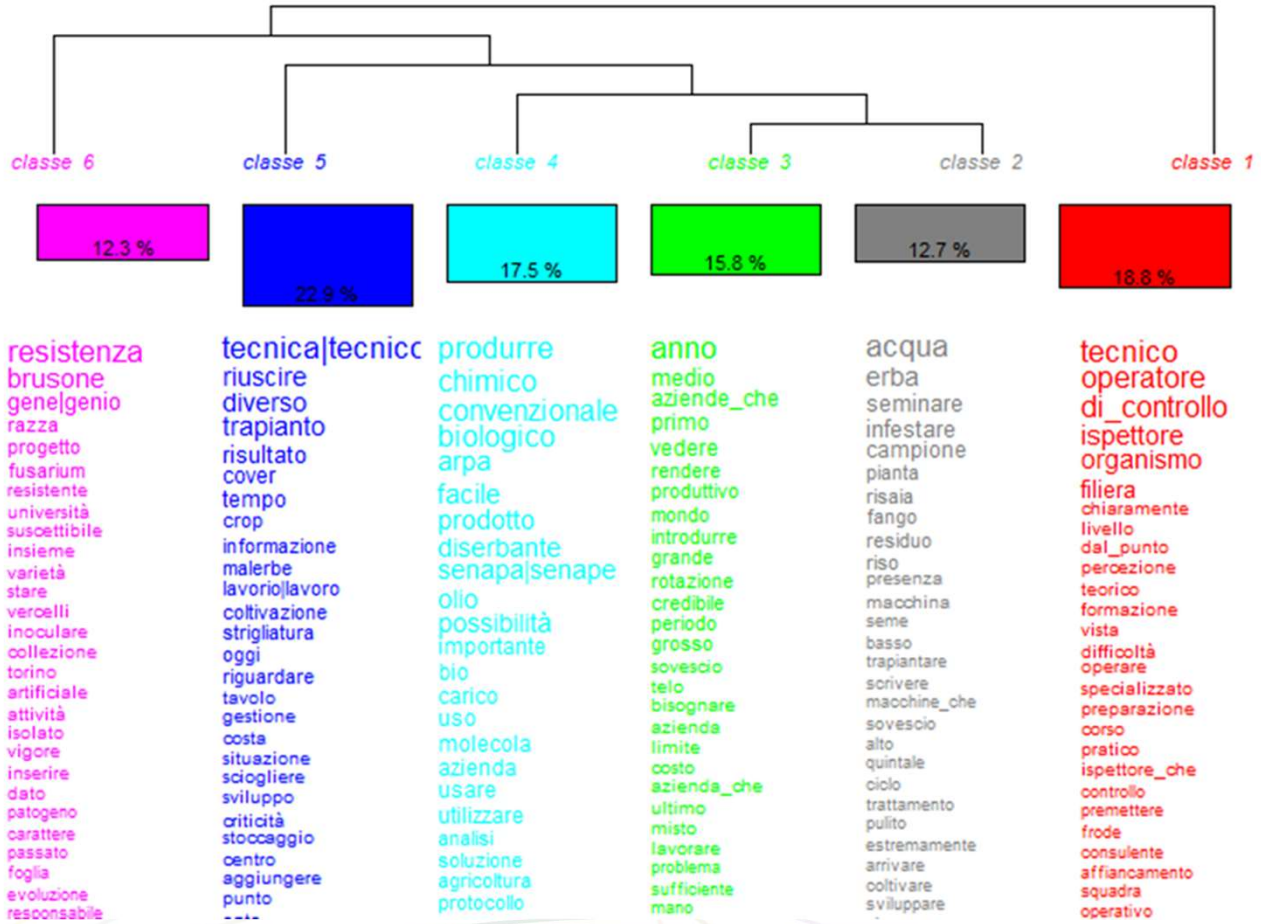


Guccione GD, Vaccaro A., Borsotto P., Vassallo M., Iacono R., Borri I. (2019). Risobiosystems: Documento di analisi del sistema di controllo e certificazione - Report WP3 Progetto di ricerca, sviluppo e trasferimento a sostegno della risicoltura biologica

Potenziali applicazioni: Focus Group – analisi trascrizioni

Riso Biologico

+-+-+-+-+-+-+-+-+
|i|R|a|M|u|T|e|Q|
+-+-+-+-+-+-+-+-+
Number of texts: 7
Number of text segments: 361
Number of forms: 2679
Number of occurrences: 13174
Number of lemmas: 1937
Number of active forms: 1585
Number of supplementary forms: 310
Number of active forms with a frequency >= 3: 494
Mean of forms by segment: 36.493075
Number of clusters: 6
292 segments classified on 361 (80.89%)

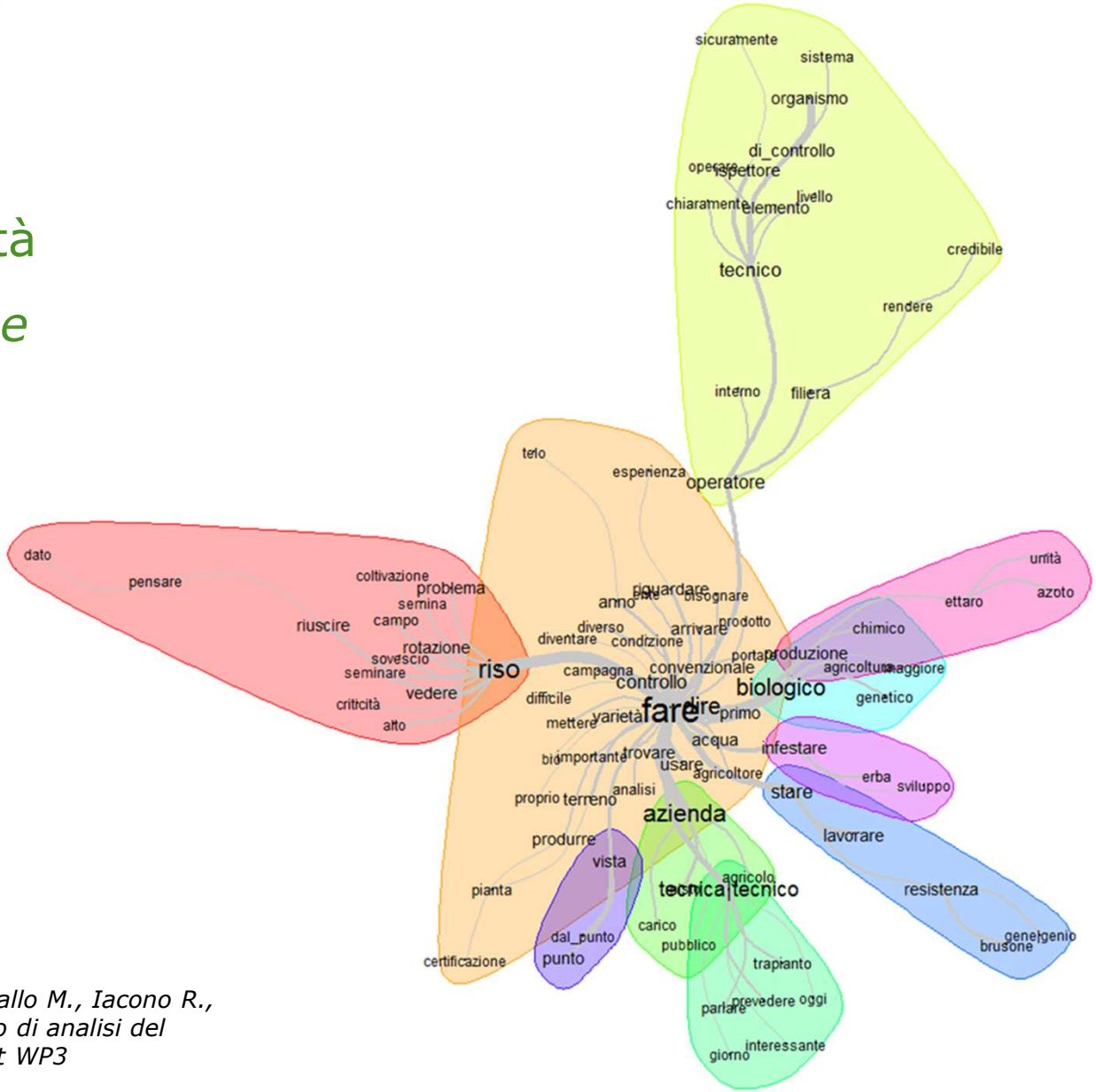


Guccione GD, Vaccaro A., Borsotto P., Vassallo M., Iacono R., Borri I. (2019). Risobiosystems: Documento di analisi del sistema di controllo e certificazione - Report WP3

Riso Biologico

Grafo delle Similarità

Co-occurrencies tree



Guccione GD, Vaccaro A., Borsotto P., Vassallo M., Iacono R., Borri I. (2019). Risobiosystems: Documento di analisi del sistema di controllo e certificazione - Report WP3

Riso Biologico

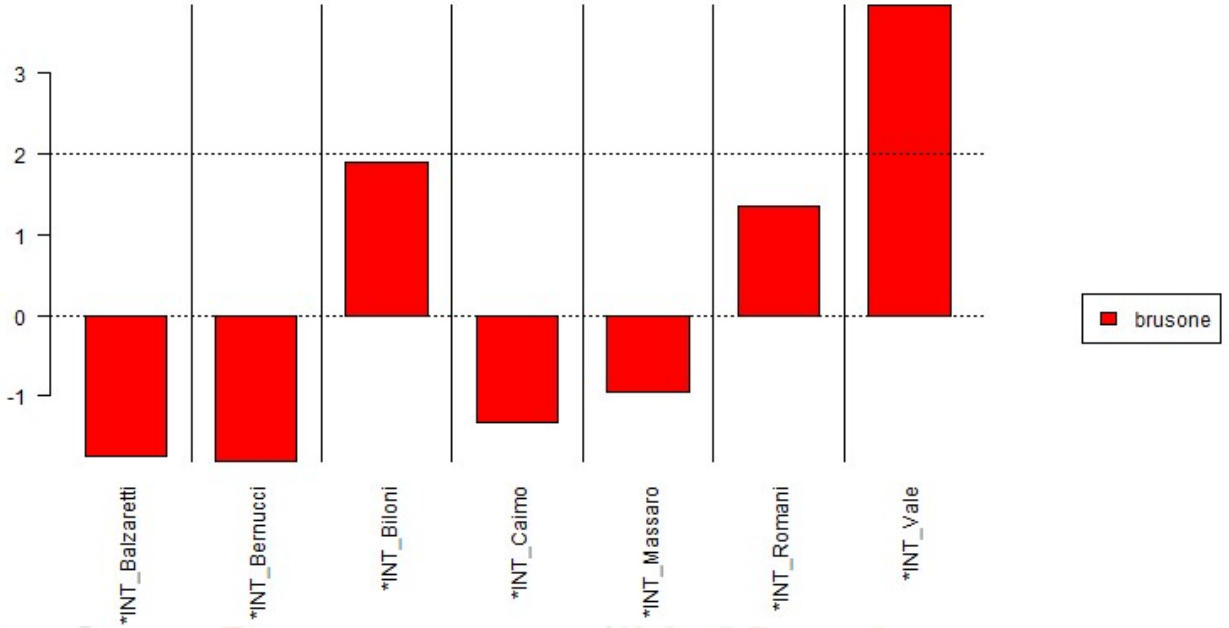
$$^{*} z_i = \frac{f_i - f_i^{rif}}{\sqrt{f_i^{rif}}}$$

$z_i > 2 \; ; caratteristica$

$z_i < -2 \; ; rara$

$z_i \approx 0 \; ; banale$

Grafico delle Specificità/Peculiarità per la parola 'brusone'



*Bolasco, S. (1999). Analisi Multidimensionale dei dati. Metodi, strategie e criteri di interpretazione. Roma, Carocci.

Guccione GD, Vaccaro A., Borsotto P., Vassallo M., Iacono R., Borri I. (2019). Risobiosystems: Documento di analisi del sistema di controllo e certificazione - Report WP3

Potenziali applicazioni: CREAgritrend

https://www.crea.gov.it/web/politiche-e-bioeconomia/-/creaagritrend

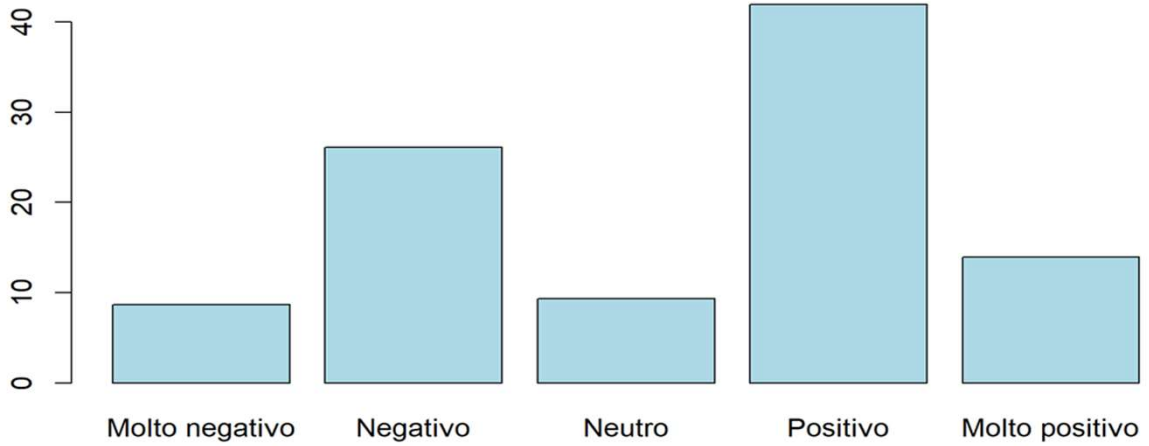
Potenziali applicazioni: CREAgritrend

PARTI SOCIALI
- Cia_Agricoltura
- coldiretti
- Confagricoltura
- UciAgricoltura
- assorurale
- AIABfederale
- Agrinsieme

MEDIA
- AgriculturaIT
- Agricolae1
- agronotizie
- AGRAPRESS
- AgricolturaNews

© Twitter

3.950 tweets, raccolti tra il 01 luglio e il 15 settembre 2019

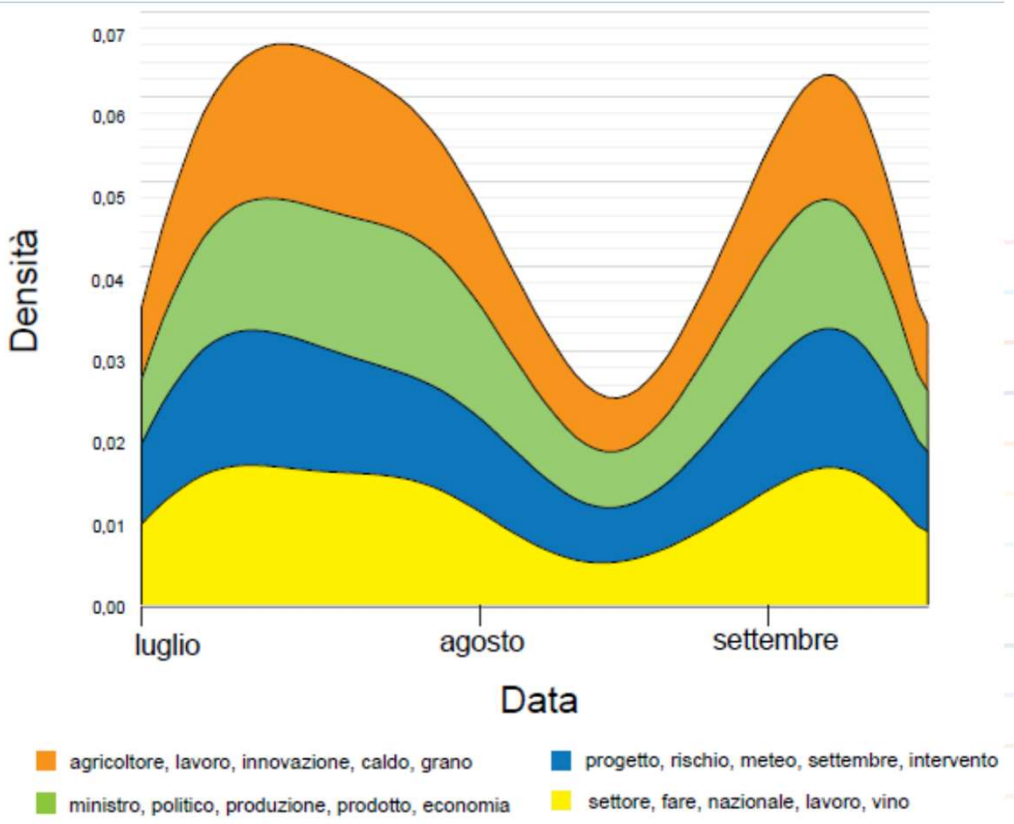


Basile V., Nissim M. (2013). Sentiment analysis on Italian tweets. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Atlanta, Georgia, USA.

Monda M., Gabrieli G., & Vassallo M. (2019). Sentiment in Agricoltura: Il termometro dell'Agricoltura. In CREAgritrend: I numeri dell'Agroalimentare Italiano. Bollettino trimestrale edito dal CREA, Centro Politiche e Bio-economia, n°3, II Semestre 2019. ISSN: 2612-6419.

Potenziali applicazioni: CREAgritrend

Hashtags word cloud



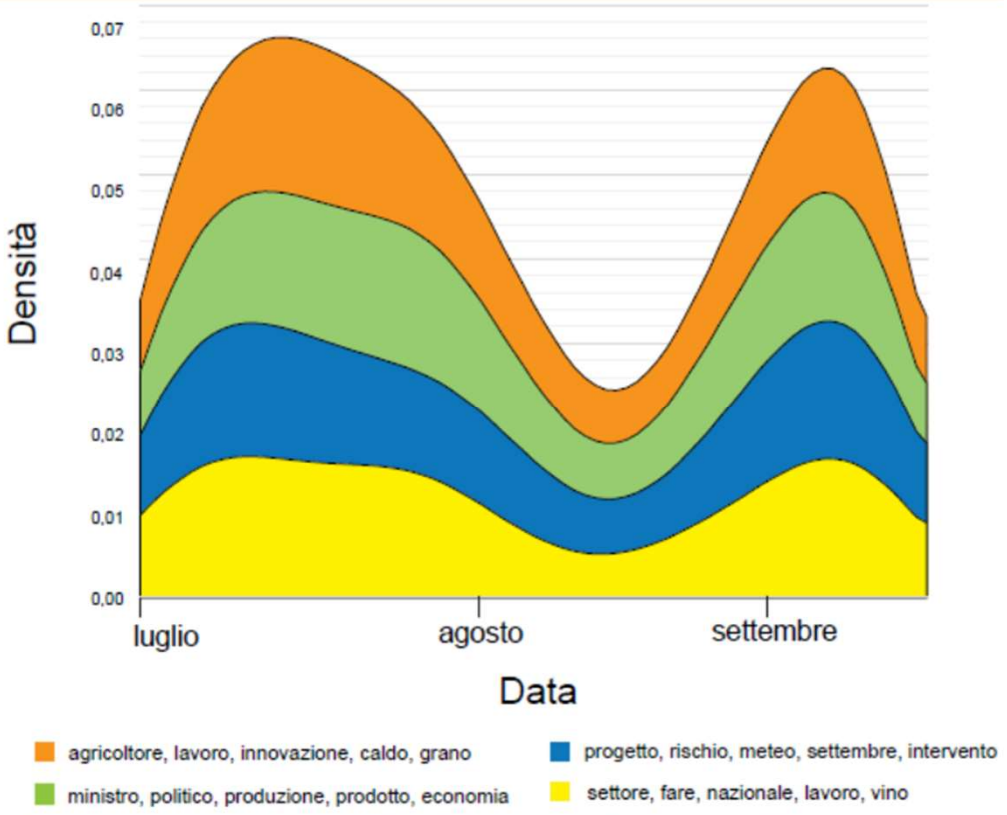
Topic Modeling

Blei, D.M. (2012). Probabilistic topic models. Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. Communication of the ACM, 55(4):77–84.

Monda M., Gabrieli G., & Vassallo M. (2019). Sentiment in Agricoltura: Il termometro dell'Agricoltura. In CREAgritrend: I numeri dell'Agroalimentare Italiano. Bollettino trimestrale edito dal CREA, Centro Politiche e Bio-economia, n°3, II Semestre 2019. ISSN: 2612-6419.

Potenziali applicazioni: CREAgritrend

- Problema Bayesiano:
  - z è un topic al quale è associata una certa probabilità $Pr(\theta_z)$ a priori (pre-sperimentale) che si verifica z
  - $Pr(\theta_z > \delta)$; probabilità che il topic sia rilevante
  - Dati osservati w(ords) relativi a z
  - $Pr(\theta_z > \delta|w)$ probabilità a posteriori
- $Pr(\theta_{z>}\delta|w) = Pr(w|\theta_z)\, Pr(\theta_z)$
  - Quindi all'estrazione di nuovi tweets (words) verifico, attraverso le probabilità condizionate a questi nuovi tweets, se il topic z è ancora rilevante
  - Le statistiche bayesiane apprendono dall'esperienza, si aggiornano ad ogni nuova immissione di dati



Topic Modeling

Blei, D.M. (2012). Probabilistic topic models. Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. Communication of the ACM, 55(4):77–84.

Monda M., Gabrieli G., & Vassallo M. (2019). Sentiment in Agricoltura: Il termometro dell'Agricoltura. In CREAgritrend: I numeri dell'Agroalimentare Italiano. Bollettino trimestrale edito dal CREA, Centro Politiche e Bio-economia, n°3, II Semestre 2019. ISSN: 2612-6419.

- Partecipazione alla «Sesta Conferenza Italiana di Linguistica Computazionale» – Bari, 13-15 /11/2019-

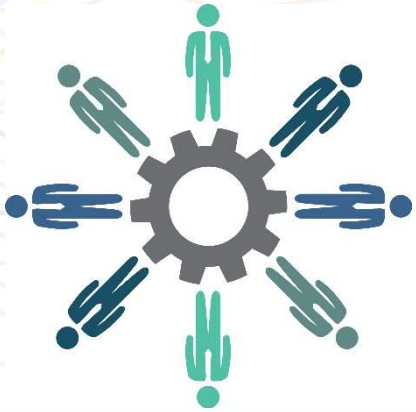
CLiC it 2019 http://clic2019.di.uniba.it/

AILC Associazione Italiana di Linguistica Computazionale http://www.ai-lc.it/

- Data set AGRITREND – tweets da Gennaio-Aprile 2019
- Problema della lemmatizzazione:
  - *Vassallo M., Gabrieli G., Basile V., Bosco C. (2019). The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis. Paper presented at the 6th Italian Conference on Computational Linguistics, Bari, November 13-15, 2019. ISSN: 1613-0073. http://ceur-ws.org/Vol-2481/paper74.pdf*
- Concetto di formario affettivo – *Morphologically-inflected Affective Lexicon (MAL)*

Prospettive future: SENTIMENT-RICA

- **Descrizione:** indagine sperimentale, a costo zero, volta a identificare e quantificare le opinioni degli imprenditori agricoli italiani che parteciperanno all'indagine campionaria **RICA*** 2019 (11.100 imprenditori) sulle politiche agricole comunitarie nel loro diretto contesto territoriale e di comparto.
- **Obiettivi:**
  - Identificare e quantificare in termini di orientamento positivo, negativo, neutrale le opinioni degli imprenditori agricoli RICA sulle politiche agricole comunitarie nel loro diretto contesto territoriale e di comparto.
  - Individuare diverse prospettive, nuovi punti di vista e valutazioni sulle politiche agricole comunitarie espresse direttamente dagli imprenditori agricoli RICA.
- **Modalità:** invio tramite PEC di due/tre domande sulla PAC pre e post 2020 senza limiti di testo nelle risposte
- **Analisi dei dati:** Text & Opinion Mining



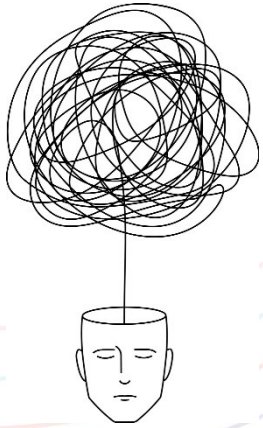
*Rete di Informazione Contabile Agricola
https://rica.crea.gov.it/index.php

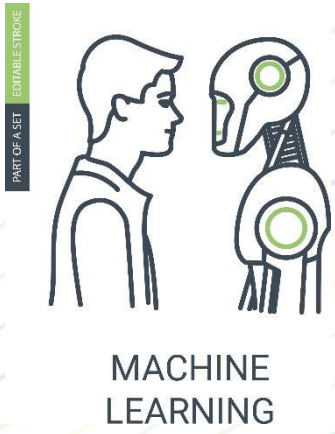
Prospettive future

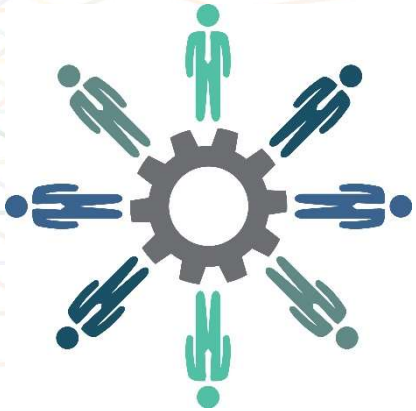
- **SOCIAL MOOD ECONOMY INDEX:** UNA MISURA DEL SENTIMENT ITALIANO SULL'ECONOMIA BASATA SUI DATI DI TWITTER
  - «L'Istat aggiorna il Social Mood on Economy Index, un nuovo indice sperimentale reso disponibile a ottobre 2018. L'indice fornisce misure giornaliere del sentiment italiano sull'economia, derivate da campioni di tweet pubblici in lingua italiana, catturati in streaming»
  - «La procedura analizza, con tecniche di sentiment analysis, una media di circa 55.000 tweet al giorno»

https://www.istat.it/it/archivio/219585

- Modelli previsionali?







Letture consigliate sull'analisi automatica dei testi

- Bolasco, S. (1999). *Analisi Multidimensionale dei dati. Metodi, strategie e criteri di interpretazione*. Roma, Carocci.
- Bolasco, S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma, Carocci.
- Giuliano, L. & La Rocca, G. (2008). *L'analisi automatica e semi-automatica dei dati testuali*. Software e istruzioni per l'uso. Milano, LED.
- Marr, B. (2015). Big Data: 20 mind boggling facts everyone must read. Forbes (online) 30 settembre. Disponibile su <https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#136378ad17b1>. [Data di accesso: 18 febbraio 2020].
- Tuzzi, A. (2003). *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Roma, Carocci.

Grazie per l'attenzione!