Innovazione: buzzword o realtà?

Baraldini Chiara, 1205133

Colombo Margherita, 1242302

Dargenio Elisabetta, 1236614

1 Introduzione

In una realtà mondiale sconvolta dallo scoppio della pandemia di covid-19, l'enorme portata dei cambiamenti che stiamo vivendo è enfatizzata da frasi come "niente sarà più come prima" e dal concetto di "un nuovo mondo" post Covid-19.
E' indubbio, infatti, che questa epidemia abbia messo in discussione lo status quo e abbia improvvisamente alterato tutti quegli equilibri geopolitici, economici, sociali e anche familiari che si erano creati nel tempo.
La diffusione del Covid-19 e le conseguenti procedure di arginamento del contagio ci hanno portati a vivere la quotidianità in un modo completamente diverso.

Anche l'Italia non è rimasta immune a questi cambiamenti o meglio, a queste opportunità di innovazione, che le nuove esigenze sanitarie, sociali e individuali hanno stimolato.

Per fare qualche esempio, le scuole hanno scoperto che si può fare didattica anche a distanza, molte aziende hanno iniziato a incentivare lo smart working, abbiamo assistito al boom dell'e-commerce e del delivery, la sanità non era mai stata così al centro dell'attenzione e la mobilità non è mai cambiata così velocemente.

Più in generale, da una parte ci troviamo di fronte a nuovi prodotti, nuovi processi e nuovi servizi o, se già esistenti, a un loro rinnovato e più intenso utilizzo, dall'altra stiamo osservando un radicale rinnovamento di interi settori.
Tutto ciò è racchiuso in una parola: innovazione.

2 Obiettivi

Questo progetto cerca di rispondere ai seguenti quesiti:

1. In quali ambiti si parla di innovazione in questo scenario post covid19? Ci si aspetta che emergano ambiti quali ad esempio l'istruzione, la sanità, l'e-commerce, le imprese, il digitale, etc.
2. Si può affermare che il covid19 abbia stimolato l'innovazione? Ovvero, gli ambiti in cui se ne parla sono cambiati rispetto ad un anno fa?
   Osservando da vicino la realtà che ci circonda, ci si aspetterebbero delle differenze, con le analisi effettuate si cercherà di chiarire se la pandemia e le turbolenze da essa portate hanno effettivamente stimolato l'innovazione.

3 Metodi e Dati

Il social media scelto per l'analisi è Twitter.
I dati sono stati estratti nel periodo che va dal 17 al 25 Maggio 2020, periodo in piena fase 2, e si è deciso di usare lo stesso identico periodo anche per il 2019, in modo da avere un confronto diretto dei due anni.
Per scaricare i dati dell'anno corrente si è utilizzata la funzione `rtweet::search_tweets()` in R, con parametro opzionale aggiuntivo `lang:it` e con numerosità `n=3000`. Dato che la funzione dà come risultato solo i tweet degli ultimi 9 giorni, ne sono stati estratti 2528.
Per i dati dell'anno passato si è utilizzato il tool `twint` in Python con parametro opzionale `lang-it`. I tweet estratti per il 2019 sono 1018.

Token.
Per l'analisi dei token si effettua una prima pulizia dei tweet usando la funzione `TextWiller::normalizzaTesti()` e togliendo caratteri speciali tramite la funzione `gsub()`. Una prima estrazione dei singoli token ha permesso di ripulire ulteriormente i tweet da parole molto ricorrenti, ma non utili ai fini dell'analisi. Vengono dunque rimossi per i tweet di entrambi gli anni i token *'wwwurlwww'*, (la stringa che la funzione `normalizzaTesti` utilizza per identificare la presenza di un link all'interno del tweet), *'più'* e *'non'*, e solo per i tweet del 2019, *'pic'*, *'twitter'*, *'com'* (*'pic.twitter.com'* indica la presenza di un'immagine all'interno del tweet).
Successivamente vengono rilevati alcuni bug nel funzionamento di questi tools, infatti tra i token più ricorrenti risultavano hashtag e mention che la funzione `search_tweets` non aveva riconosciuto come tali (ad esempio *'bastaprecariatodistato'* e *'mise_gov'*). Per risolvere questo problema si decide di eliminare dal testo tutti i token con il primo carattere uguale a '#' o '@'.

```
dati$text<- gsub('#\\S+', '', dati$text)
dati$text<- gsub('@\\S+', '', dati$text)
```

L'estrazione dei token (singoli e bigrammi) è stata effettuata con `tidytext::unnest_tokens()`.

Hashtag.
Si decide inoltre di fare un'analisi anche sugli hashtag, poiché potrebbero risultare più rappresentativi dei temi trattati all'interno dei tweet rispetto alle semplici parole.
L'analisi degli hashtag ha richiesto alcune manipolazioni iniziali dei dati per ottenere le strutture richieste nei pacchetti `dplyr` e `tidyr`.
Relativamente ai dati del 2020, come già anticipato, alcuni hashtag non venivano riconosciuti come tali, e la variabile *hashtag* del tibble comprendente i dati in esame, conteneva solo una parte di quelli presenti nei relativi tweet. Per ovviare a questo problema, tramite Python, sono stati estratti i singoli token dai tweet contenenti ancora hashtag e mention, sono stati salvati in una lista di liste (ogni sottolista contiene i token relativi ad un singolo tweet), ed infine sono stati salvati solo quelli che iniziavano con "#" . Questa lista di liste è stata poi caricata su R e modificata come di seguito:

```
hashtag_innovazione_py <- read.csv('hashtag_innovazione20_postpy.csv',
                                   sep='\n',header = F,blank.lines.skip = F)
lista_hashtag<-list()
for (i in (1:dim(hashtag_innovazione_py)[1])){
  lista_hashtag[i]<- as.list(strsplit(hashtag_innovazione_py[i,], ","))
  }

matrice <- (matrix(lista_hashtag,nrow = 2528))
df_hash_20 <-as.data.frame(matrice)
```

In questo modo la nuova variabile *df_hash* contenente gli hashtag ha la stessa struttura della variabile *hashtag* originale, e infine si utilizzano i tools dei pacchetti `dplyr` e `tidyr` per separare e contare gli hashtag.

Per i dati del 2019 invece la variabile *hashtag* conteneva già tutti gli hashtag utilizzati nei relativi tweet, in questo caso però la struttura della variabile era diversa dai dati del 2020, poiché la procedura di estrazione dati utilizzata era stata differente.
Perciò anche in questo caso sono state necessarie delle manipolazioni, in modo da poter utilizzare gli stessi tool usati per i dati del 2020.

```
hashtag19<-dati19$hashtags
lista_hashtag_19=list()
for (i in (1:1018)){
  hashtag19[i]<-gsub("\\[|]","",hashtag19[i])
  hashtag19[i]<-gsub("\\'","",hashtag19[i])
  hashtag19[i] <- str_split(hashtag19[i],", ")
  lista_hashtag_19[i]<-hashtag19[i]
}

matrice <- matrix(lista_hashtag_19, nrow = length(lista_hashtag_19))
df_hash_19 <-as.data.frame(matrice)
```

Anche in questo caso si utilizzano i tools dei pacchetti `dplyr` e `tidyr` per separare e contare gli hashtag.

LDA.
Si decide di effettuare sui dati estratti anche un'analisi *Latent Dirichlet Allocation* (LDA), il principale modello basato sugli argomenti, in modo da valutare se emergono argomenti differenti nei dati del 2020 (sanità, istruzione, …), e se emergono differenze tra i dati dello scorso anno rispetto a quello corrente.
Inizialmente si trasformano i dati in formato `documentTermMatrix`, del pacchetto `tm`. Questo formato è quello richiesto per applicare la *Latent Dirichlet Allocation*.

```
tt_ww<- dati%>%select(text,id)%>%
  unnest_tokens(word, text)%>%
  group_by(id,word)%>%
  mutate(freq=n())
tt_ww
dtm<-tt_ww%>% cast_dtm(document=id, term=word, value = freq)
```

Il file dtm così ottenuto ha dimensione:

```
> dim(dtm)
[1] 2518 1131
```

Si decide di concentrarsi solo sui termini che compaiono almeno in 5 documenti. Per questa operazione si utilizza la funzione `removeSparseTerms()`.
Si rimuovono infine i documenti rimasti vuoti a causa dell'ultima operazione di pulizia.

```
tresh<-1-5/dim(dtm)[1]
dtm<-removeSparseTerms(dtm,tresh)
ui<-unique(dtm$i)
dtm.new<-dtm[ui,]
```

La dimensione del nuovo oggetto dtm è dunque:

```
> dim(dtm.new)
[1] 2498 1131
```

Si procede dunque alla stima del modello LDA, attraverso la funzione `LDA()` all'interno del pacchetto `Topicmodels`, utilizzando come parametro `method='Gibbs'`: è uno dei metodi che permettono di costruire algoritmi per simulare valori da densità multivariate complesse, in particolare in questo caso, simulano i valori della distribuzione a posteriori $\pi(\beta, \gamma \mid w)$, dove β_k=distribuzione dei termini nel topic k e γ_i=distribuzione dei topic nel documento i.
Dopo alcuni tentativi, si decide di stimare il modello con 4 topic.

```
library(topicmodels)
q_lda<- LDA(dtm.new,k=4, method = 'Gibbs',
            control=list(seed=1))
```

Si procede analizzando le stime dei parametri tramite media a posteriori. Si estrae la matrice di coefficienti β_k e si ordinano i termini in base ai valori β_{kv}, per identificare all'interno di ogni topic quali sono i termini con probabilità maggiori.

```
beta_topics<-tidy(q_lda, matrix='beta')

termini<-beta_topics%>% group_by(topic)%>%
  top_n(10,beta)%>% arrange(topic)
```

Si estraggono poi i coefficienti $\gamma_{i,}$ per avere la distribuzione dei topic all'interno di ogni tweet.

```
gamma_doc<- tidy(q_lda, matrix='gamma')
```

Si valuta per ogni tweet qual è il topic al quale viene assegnata una probabilità a posteriori maggiore.

```
doc_best<-gamma_doc%>% group_by(document)%>%
  top_n(1,gamma)
doc_best%>% ungroup()%>% arrange(desc(gamma))
```

Si estraggono, per ogni topic, i tweet con valori massimi di γi, e che possiamo interpretare come quelli meglio classificati.

```
doc_id<-gamma_doc%>% group_by(topic)%>%
  arrange(gamma)%>%mutate(ord=1:n())

doc_best<-doc_id%>% filter(ord>=n())
```

Infine si utilizza la funzione `right_join()` per recuperare i tweet nel dataset originale.

```
id_best<-doc_best%>% mutate(id=as.numeric(document))

res<-innovazione%>% select(screen_name, text, id)%>%
  right_join(id_best, by='id')
```

4 Risultati

I 20 **token** con maggiore frequenza assoluta estratti per i dati del 2020 e del 2019 risultano essere rispettivamente:

	word	tot
	<chr>	<int>
1	oggi	227
2	presidente	165
3	paese	157
4	nuovo	148
5	post	133
6	ricerca	120
7	speciale	117
8	futuro	113
9	motore	112
10	ancora	108
11	opportunità	108
12	grazie	106
13	2020	105
14	aziende	104
15	qui	103
16	investire	96
17	essere	94
18	fondamentale	93
19	maggio	93
20	servizi	93

	word	tot
	<chr>	<int>
1	oggi	71
2	grazie	50
3	progetto	50
4	scopri	48
5	maggio	40
6	italia	38
7	imprese	37
8	sempre	37
9	2019	36
10	sviluppo	36
11	digitale	35
12	fare	32
13	futuro	32
14	qui	32
15	evento	30
16	mondo	30
17	essere	28
18	giugno	27
19	lavoro	27
20	ricerca	27

Dal confronto di questi output emergono alcuni aspetti interessanti.
Innanzitutto, la parola che ricorre con maggiormente sia nei dati del 2019 che in quelli del 2020 è la stessa, ovvero “oggi”. Tuttavia, la sua frequenza di utilizzo da un anno all’altro è più che triplicata (71 occorrenze nel 2019, 227 nel 2020).
In generale, si nota che le frequenze dei singoli token per l’anno 2019 sono di molto inferiori rispetto a quelle del 2020. Questo deriva naturalmente dalle numerosità dei tweet estratti per i due anni: a parità di arco temporale, i tweet contenenti l’hashtag innovazione nel 2020 sono più del doppio di quelli nel 2019.

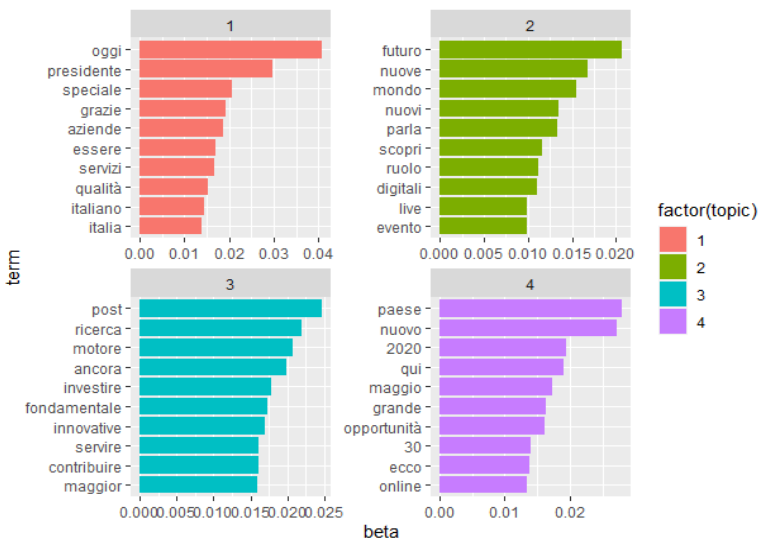
Nel 2020 emerge tra le prime parole “post”, che verosimilmente si potrebbe riferire a “post covid19”. La parola “ricerca” emerge in entrambi gli anni, ma mentre nel 2019 è la ventesima parola, nel 2020 è la sesta parola; questo rispecchia quanto in questo periodo in particolare sia stia puntando sulla ricerca per produrre innovazione. In generale, dall’output relativo al 2020 l’innovazione appare come “fondamentale” e viene vista come “motore” del cambiamento per sfruttare le nuove “opportunità” per le aziende e per il Paese, inoltre emerge anche il bisogno di “investire” di più in innovazione.
Nel 2019, invece, le parole più usate rimandano a un’idea di innovazione legata principalmente alle “imprese”, al “digitale” e al “lavoro”, come ci si aspettava.

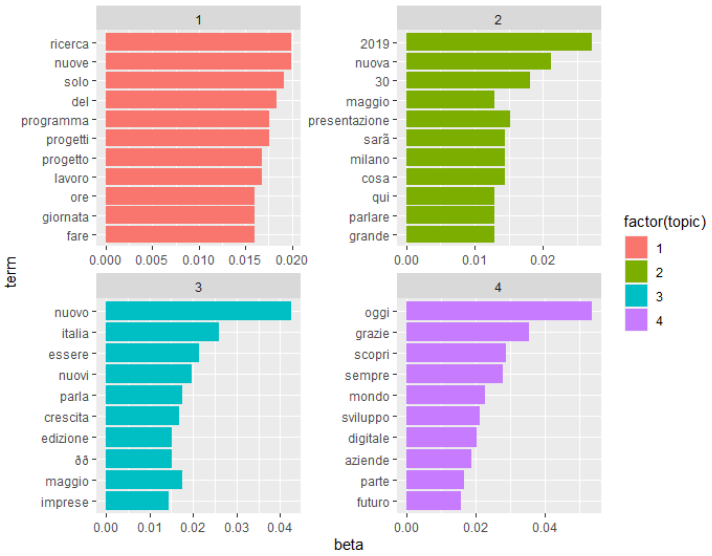
Per quanto riguarda l’analisi sugli **hashtag** invece, quelli con frequenza più elevata per 2020 e 2019 sono rispettivamente:

Si nota come a predominare nei tweet selezionati del 2020 siano hashtag riguardanti la situazione attuale, infatti troviamo “covid19”, “covid_19” e “coronavirus”. Anche gli hashtag relativi alla ricerca sono tra i più utilizzati, infatti emergono “ricerca” “precari” e “bastaprecariatodistato”, quest’ultimo è infatti un hashtag parte di un movimento che si è mosso in questi ultimi mesi contro il precariato nella ricerca pubblica. Tra gli altri si trova anche “smartworking”, altro fenomeno che ha avuto uno sviluppo enorme negli ultimi mesi, soprattutto durante fase 1 e fase 2.
Contrariamente a quanto ci si aspettasse, questi risultati non evidenziano eterogeneità negli argomenti trattati, ma sembra esserci una polarizzazione verso il tema pandemia/ripartenza.
Si potrebbe approfondire l’indagine rimuovendo gli hashtag relativi a covid19 per verificare se escludendo questo tema emergono distinzioni tra ambiti.

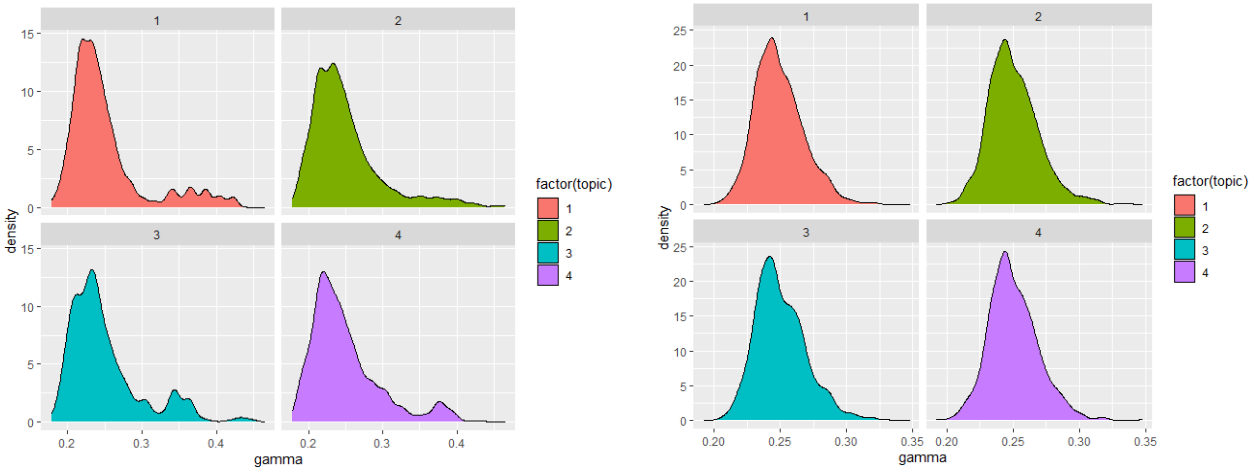
In entrambi gli anni, l'hashtag che per eccellenza viene utilizzato insieme alla parola innovazione risulta "startup"; anche questo è un risultato che non stupisce, dal momento che per loro natura le startup hanno un grande potenziale innovativo, nel passato come nel presente.
La differenza principale tra i due anni che si può cogliere è che nel 2019 gli argomenti legati all'innovazione erano limitati al mondo delle imprese (sono presenti infatti gli hashatag "imprese", "business", "pmi", "lavoro") e affini (dalla "digital transformation" all'"industria 4.0", passando attraverso la "blockchain" e l'"intelligenza artificiale"/ "ai").
Dai tweet estratti dal 2020 invece, sebbene non sia evidente una chiara separazione di tematiche, si può tuttavia notare un ampliamento del raggio; in altre parole, quando si parla di innovazione, il focus non è più verticale e in un certo senso tecnico, ma si sviluppa su un piano principalmente orizzontale, spaziando attraverso l'attualità.

Nell'analisi **LDA**, vengono estratti i beta per ogni topic, e si costruisce il grafico per vedere all'interno di ogni topic quali sono le parole più identificative.
I grafici per gli anni 2020 e 2019 sono rispettivamente:





Vengono estratti anche i gamma, e tramite il grafico si vede come si distribuiscono le probabilità tra i diversi topic.



I tweet meglio classificati per l'anno 2020 dei topic 1, 4, 3, 2 sono rispettivamente:

```
[1] "orgoglioso qualità presidente @angi_tech intervenire oggi pa social day grazie particolare amici #pas
ocial francesco costanzo christian tosolin # #digitale #business #comunicazione #giovani "

[2] "ecco whitehall machine learning with knime challenge lunedì 25 maggio sfideremo tema #machinelearning
 attraverso piattaforma #knime # #challenge #lifeatreply clicca qui maggiori informazioni "

[3] " ora emergenza stata acceleratore modernizzazione tornare indietro impensabile # nutrita discontinuit
à serve unq visione politiche industriali medio lungo termine riprogettare #lavoro @ldraimondo @sapienzaro
ma "
[4] "studente interessato #ricerca #policymaking vuoi lavorare esperti campi # #finanza politica economica
 affari europei action institute offre opportunità #stage invia cv cover letter recruiting @actioninstitut
e org #takeaction "
```

I tweet meglio classificati per l'anno 2019 dei topic 4, 2, 3, 1 sono rispettivamente:

```
[1] "#artpiub2019 day programma ultimo giorno talk show artathon talk arte economia mostre aperte fino 20 00 "

[2] "già lavorando salone #formazione #innovazione musicale 2020 convinti prossima edizione aggiungeranno ancora pià scuole
 ancora pià aziende importanti professionisti settore decisi progettare insieme futuro #musica pic twitter com e8xmy7guk3"
[3] "domani 24 05 ore 30 @confindustriasa svolgerà prima tappa progetto #innovationroadlab finalizzato stimolare confronto
 istituzioni imprenditori manager apportare #innovazione azienda iscriversi clicca qui "
[4] "alberto mattiello keynote speaker assemblea piccola #industria aib parleremo #innovazione #pmi evento programma 30 05
 2019 15 30 gratuito aperto previa iscrizione "
```

2020

Dai topic emersi dal 2020 non si nota una netta distinzione degli argomenti, tuttavia è possibile individuare alcune caratteristiche che li differenziano. Questo ci viene confermato anche dal grafico della distribuzione dei gamma, le curve infatti non sono perfettamente identiche, ma presentano delle variazioni sui valori più alti di gamma.
Inoltre va ricordato che anche se la parola Covid non è emersa in nessuno dei topic, -probabilmente perché utilizzata più come hashtag più che come token-, questo costituisce un elemento intrinsecamente comune a ciascun raggruppamento.

Il topic 1 appare più focalizzato sulla situazione attuale dell'Italia e delle aziende italiane, infatti tra le 10 parole che più rappresentano questo raggruppamento troviamo “oggi”, “Italia”, “italiano”, “paese”, “presidente”, “aziende”, “servizi”.

Il topic 2 potrebbe trattare invece dell'innovazione in ottica più globale e con un orientamento al futuro e percezione di un nuovo mondo post Covid, si notano infatti le parole “futuro”, “nuove”, “nuovi”, “mondo”; si parla anche del crescente ruolo del digitale anche come luogo di eventi (“ruolo”, “digitali”, “live”, “evento”)

Il topic 3 è incentrato sull'innovazione come chiave per la ripartenza e si nota chiaramente da parole come "post (Covid)"," ricerca”, “motore”, “investire”, “fondamentale”, “innovative". Il tweet che emerge come quello classificato in modo meno incerto di questo topic, è perfettamente in linea con quanto appena detto, poiché parla della riprogettazione della ripartenza.
Il topic 4 è un po' più difficile da inquadrare, tuttavia guardando alle parole “paese”, “nuovo”, “opportunità” e “online” si può dire che questo raggruppamento potrebbe sottolineare le nuove opportunità aperte dagli eventi online in un periodo in cui quelli offline non erano consentiti. Anche il tweet “più rappresentativo” di questo topic conferma questa interpretazione, poiché parla dell’utilizzo di una piattaforme online per il lancio di una challenge.

2019

Nel 2019 la separazione tra i 4 gruppi è ancora meno chiara e netta, dal momento che tutti trattano dell'innovazione all'interno del mondo del lavoro e delle imprese. Infatti le probabilità dei gamma sono molto simili e centrate sul valore 0.25.

Nel topic 1 si parla di un tipo di innovazione che potremmo definire programmata, infatti vediamo parole come "ricerca", "programma", "progetti", "progetto", "lavoro", "ore", "giornata". Il tweet più rappresentativo di questo raggruppamento ci aiuta a capire che la parola “programma” potrebbe essere intesa come scaletta organizzativa di un evento proposto.

Nel topic 2 cambia la prospettiva e l'innovazione si ritrova all'interno di grandi eventi e questo si può notare da parole come "presentazione", "Milano", "grande", "parlare".

Il topic 3 è più focalizzato sull'Italia, le sue imprese e sulla loro crescita, infatti il tweet più rappresentativo parla di un progetto innovativo di Confindustria per stimolare il confronto tra istituzioni, imprenditori e manager d’azienda.

Il topic 4 torna a trattare dell'innovazione ma con un taglio più globale, si parla di aziende e del digitale (con parole come "mondo", "sviluppo", "digitale", "aziende"). Questo raggruppamento è concentrato sul presente ("oggi" è infatti la parola più frequente"), ma non tralascia di parlare del futuro ("futuro" è in decima posizione).

5 Conclusioni

Dalle analisi effettuate emergono due considerazioni principali.

La prima osservazione che si può fare alla luce dei risultati ottenuti è che, nel periodo temporale considerato nel 2020 rispetto allo stesso periodo dell'anno prima, si parla molto di più di innovazione (circa 2.5 volte di più).
Questo conferma quanto ci si attendeva: il coronavirus ha effettivamente stimolato l'innovazione o, quantomeno, il dibattito intorno ad essa.

La seconda cosa che si può notare osservando quanto emerge dalle varie analisi è che l'innovazione, per quanto in questo periodo sia stimolata dal coronavirus, non è riconducibile ad alcun settore in particolare; pare, al contrario, che essa sia un fenomeno generale contrariamente a quanto si era ipotizzato inizialmente.

Limiti del progetto: c'è una possibilità che non siano emerse distinzioni significative tra gli ambiti coinvolti dall'innovazione poiché i contenuti dei tweet sono più generici rispetto, per esempio, al contenuto dei bandi governativi per l'innovazione dei servizi pubblici. È dunque possibile, che non riesca ad emergere l'eterogeneità degli ambiti che hanno subito un'ondata di innovazione in quest'ultimo periodo, perché non se ne discuta a sufficienza su Twitter.