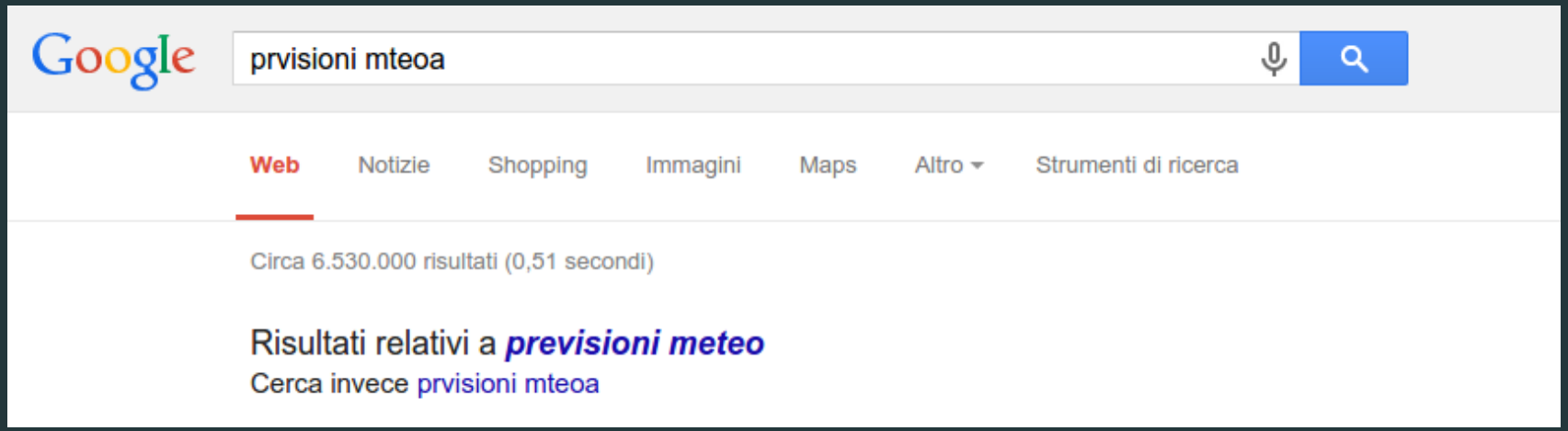
SULLA CORREZIONE ORTOGRAFICA AUTOMATICA

Un’applicazione concreta partendo da Wikipedia

Luca Chiodini

Esame di Stato 2015

ITIS Paleocapa



Cosa c'è dietro le quinte?

correzione (unigrammi)

amre

Correzione (unigrammi)

amre

Amore? Mare? Amare?

correzione (unigrammi)

amre

Amore? Mare? Amare?

$$\arg\max_{c \in D} \underbrace{P(w|c)}_{\text{error}} \cdot \underbrace{P(c)}_{\text{language}}$$

correzione (ngrammi)

nuotano nel **amre**

correzione (ngrammi)

nuotano nel **amre**

Amore? Mare? Amare?

correzione (n-grammi)

l’**amre** è cieco

Amore? Mare? Amare?

CORREZIONE (N-GRAMMI)

l'**amre** è cieco

Amore? Mare? Amare?

$$\arg\max_{c \in D} \underbrace{P(w|c_i)}_{\text{error}} \cdot \underbrace{P(c_{i-1}|c_i) \cdot P(c_i) \cdot P(c_i|c_{i+1})}_{\text{language}}$$

WIKIPEDIA

- Wikipedia in italiano come sorgente di dati

WIKIPEDIA

- Wikipedia in italiano come sorgente di dati
- Analizzare **tutta** Wikipedia **non è banale**!

WIKIPEDIA

- Wikipedia in italiano come sorgente di dati
- Analizzare **tutta** Wikipedia **non è banale**!
- ~ 10 GB di pagine (solo testo)

WIKIPEDIA

- Wikipedia in italiano come sorgente di dati
- Analizzare **tutta** Wikipedia **non è banale**!
- ~ 10 GB di pagine (solo testo)
- Elaborazione su Amazon EC2 (8 CPU, 15 GiB RAM)

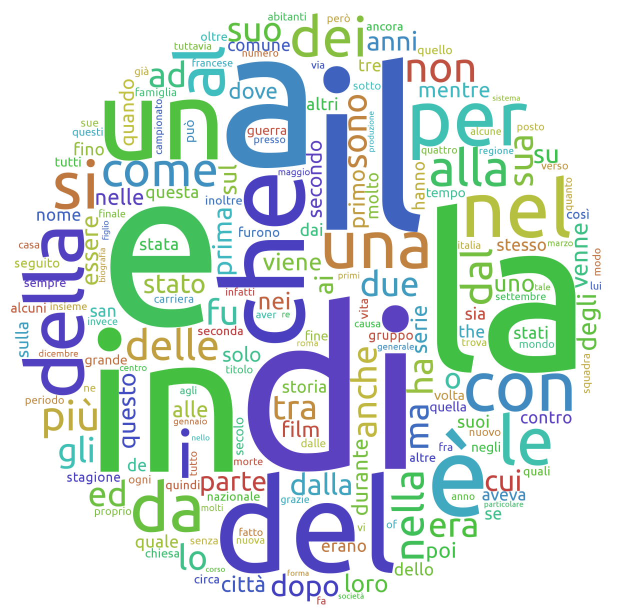
WIKIPEDIA

- Wikipedia in italiano come sorgente di dati
- Analizzare **tutta** Wikipedia **non è banale**!
- ~ 10 GB di pagine (solo testo)
- Elaborazione su Amazon EC2 (8 CPU, 15 GiB RAM)
- ~ 311 000 000 parole (~ 2 600 000 diverse)

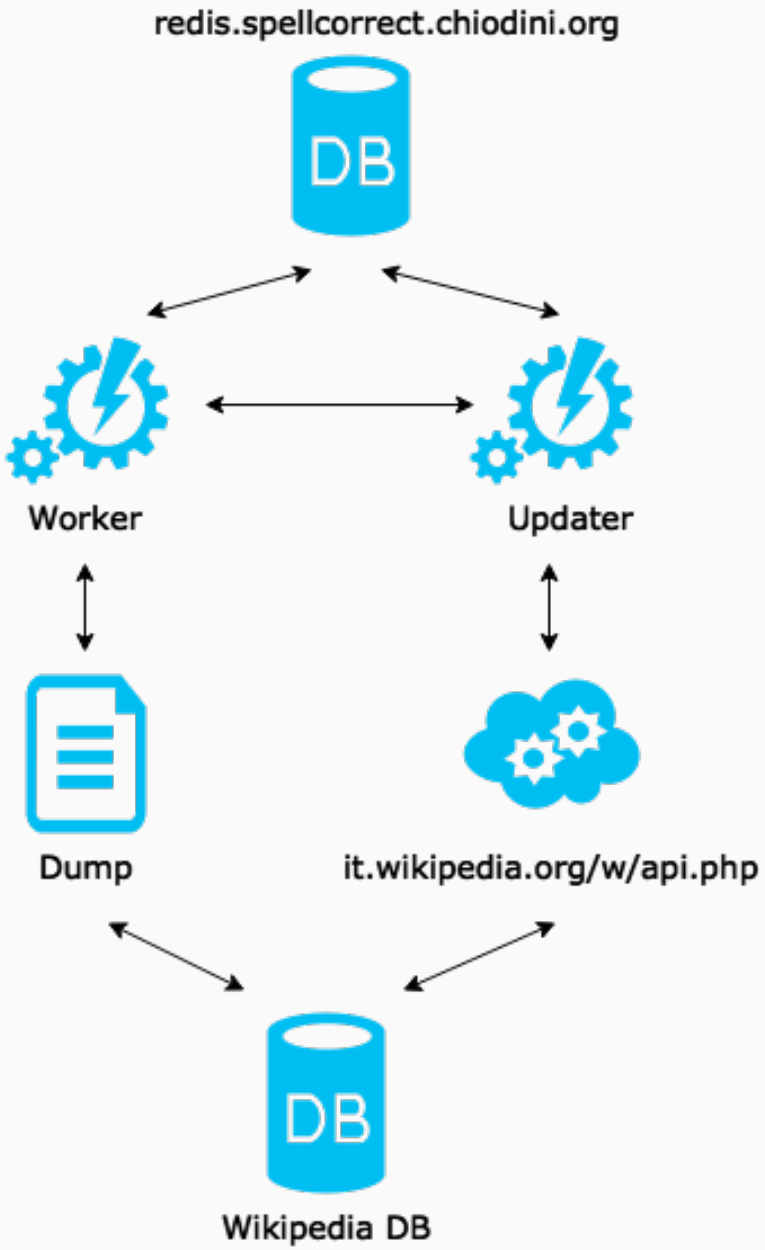
WIKIPEDIA

- Wikipedia in italiano come sorgente di dati
- Analizzare **tutta** Wikipedia **non è banale**!
- ~ 10 GB di pagine (solo testo)
- Elaborazione su Amazon EC2 (8 CPU, 15 GiB RAM)
- ~ 311 000 000 parole (~ 2 600 000 diverse)
- ~ 41 000 000 coppie di parole diverse

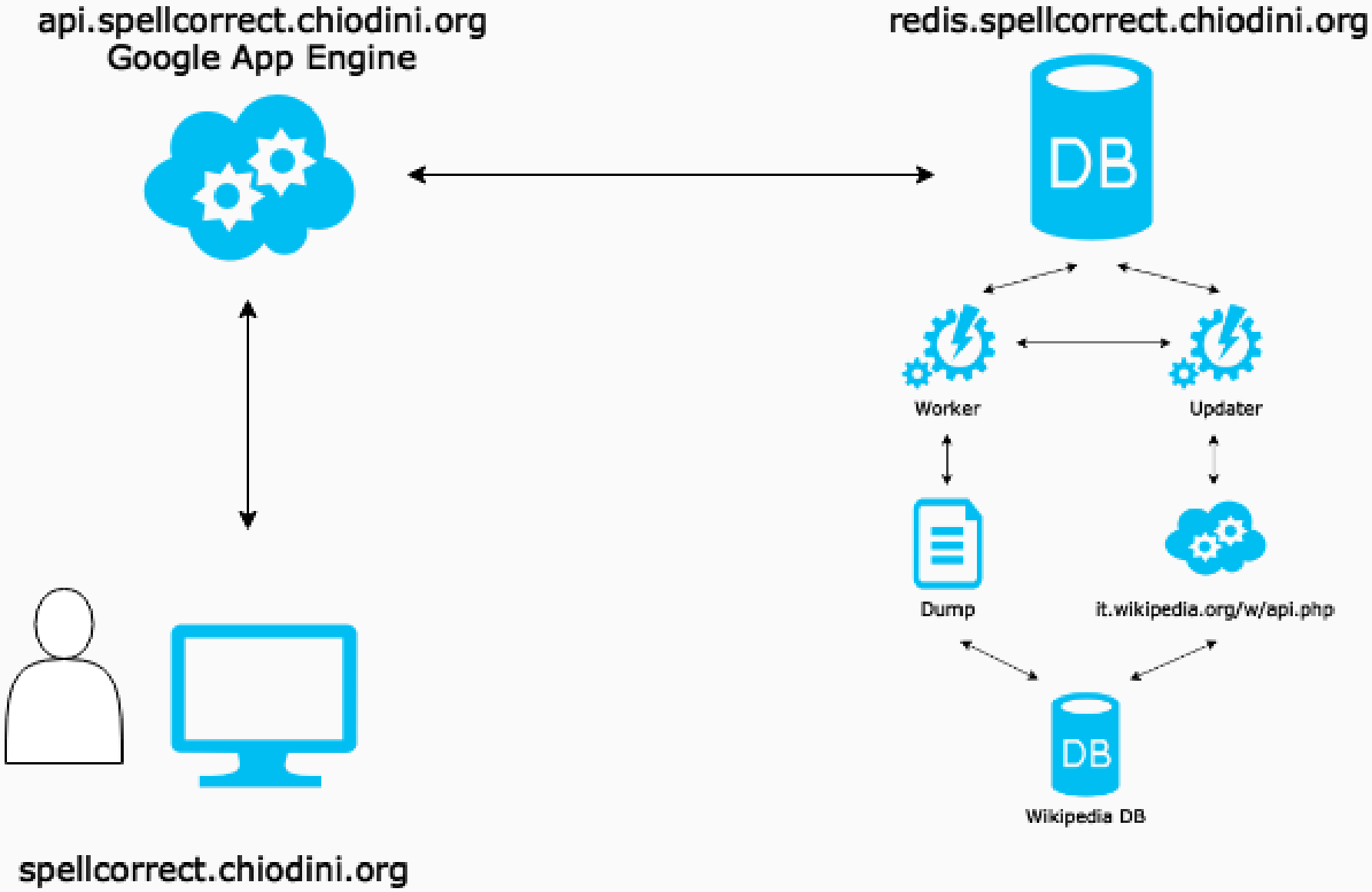
WORD CLOUD


abitanti
però
ancora
suo
dei
anni
oltre
comune
tuttavia
numero
quello
francese
via
tre
non
già
ad
al
sotto
mentre
famiglia
dove
sistema
quando
campionato
può
una
a
altri
il
per
sue
questi
guerra
presso
sono
produzione
posto
alcune
fino
quattro
regione
tutti
si
come
maggio
secondo
alla
sua
su
verso
nelle
questa
sul
inoltre
primo
molto
hanno
tempo
nome
finale
figlio
che
nel
quanto
così
casa
stata
e
furono
dai
una
italia
stesso
marzo
seguito
biografia
essere
prima
viene
primi
venne
modo
sempre
della
stato
carriera
ai
due
dal
uno
tale
settembre
lui
infatti
alcuni
insieme
san
invece
nei
aver
re
vita
la
sia
degli
fu
causa
seconda
serie
the
stati
delle
fine
roma
gruppo
trova
mondo
sulla
dicembre
solo
generale
grande
centro
titolo
anche
squadra
o
ne
storia
con
periodo
agli
tra
ha
volta
più
in
alle
di
quella
contro
questo
gennaio
film
ma
suoi
le
nuovo
secolo
i
nello
dalle
durante
fra
negli
gli
tutto
morte
altre
de
è
quali
stagione
ogni
quindi
parte
dalla
nella
cui
ed
nazionale
anno
aveva
da
molti
grazie
particolare
proprio
del
se
of
vi
era
quale
fatto
senza
nuova
erano
chiesa
poi
lo
corso
forma
dello
circa
città
dopo
loro
fa
società

ARCHITETTURA DEL SISTEMA



ARCHITETTURA DEL SISTEMA



Tutto ciò che è bello e nobile è il risultato della ragione e di calcoli.

Baudelaire