

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in Informatica

ELABORATO FINALE

DANDELION

Supervisore Laureando

Alberto Montresor Edoardo Lenzi

Anno accademico 2018

Contents

Sommario

Il progetto di tesi consiste in una prima parte riguardante l'implementazione di un client in C#, finalizzato allo scopo di includere in una libreria metodi plug&play per chiamare automaticamente le RESTful-API del servizio online Dandelion.

Durante lo sviluppo il codice é stato regolarmente caricato su GitHub per una periodica revisione da parte del team di SpazioDati; per rendere lo sviluppo piú efficiente si é cercato di seguire i principi del TDD, affidandosi al servizio online Travian per validare automaticamente ogni rilascio.

Una volta completata la libreria, é stata stilata la documentazione (esportata in html tramite il tool Wyam) e caricata la dll su Nuget; pertanto la libreria é ora facilmente integrabile in qualsiasi progetto C#.

La seconda parte del progetto di tesi riguarda sempre il servizio Dandelion di SpazioDati ma, questa volta, una parte del codice back-end. Partendo dall'implementazione attualmente in uso si sono studiate delle implementazioni alternative per massimizzare la velocitá dell'algoritmo e minimizzare la memoria occupata dalle strutture dati di appoggio dell'algoritmo.

. . .

Al termine dell'esperienza si é concluso che ...

1 Client C#

1.1 Dandelion

Dandelion é un servizio online fornito da Spazio Dati che mette a disposizione dell'utente servizi di analisi semantica testuale; grazie ad esso é possibile, dato un testo, estrarne le entitá semantiche principali (*Entity Extraction*), trovare la lingua in cui é stato scritto (*Language Detection*), classificarlo secondo modelli definiti dall'utente stesso (*Text Classification*) e analizzarne la semantica per capirne i sentimenti che l'autore ci vuole trasmettere (*Sentiment Analysis*).

Dandelion ha in oltre delle RESTful-API che permettono di confrontare due testi generando un indice di similitudine fra i due (*Text Similarity*) e un motore di ricerca di entitá di Wikipedia (*Wikisearch*), nel caso si voglia trovare il titolo di un contenuto senza conoscerlo a priori.

1.2 Struttura Generale

Per l'implementazione si é cercato di seguire le best practices, dettate dalle linee guida Microsoft, per la stesura del codice. Sono stati adottati, come pattern di programmazione, dependency injection, MVC (per quanto possibile) e TDD, appoggiandosi a librerie esterne Nuget quali Newtonsoft. Json e SimpleInjector. I metodi sono asincroni e le property thread-safe.

La solution é stata partizionata in vari progetti; il progetto Business contiene la business-logic (comprendente servizi, metodi estensione, l'implementazione del client e un wrapper del Container di SimpleInjector) e utilizza i modelli di Domain.

In Test sono contenute le classi di testing e le fixture XUnit, mentre in Documentation sono presenti i file generati dal tool Wyam per la documentazione.

1.3 Il Client

Dalla documentazione ufficale della API di Dandelion si evince che ogni end-point richiede uno o più testi ed una serie di parametri alcuni obbligatori ed altri opzionali; pertanto sono stati creati dei

modelli, coerenti con tali definizioni, che, passati in argomento a dei servizi, ritornano i DTO delle risposte degli end-point di Dandelion.

Ogni servizio controlla che i parametri inseriti dall'utente rispettino i constraints definiti nella documentazione e costruiscono tramite il metodo ContentBuilder() un dizionario di parametri e tramite UriBuilder l'URI che identifica l'end-point.

Infine il service chiama il metodo generico CallApiAsync() passandogli il content, la URI ed il metodo HTTP; nel metodo di chiamata si gestiscono separatamente i metodi HTTP Get e Delete perché non permettono il passaggio parametri nel boby pertanto è necessario fare il percent-encoding del content e passarlo come query parameter. Dandelion ammette sia chiamate in Get che in Post ai servizi tuttavia nel caso di Get non garantisce il corretto funzionamento del servizio per testi che superano i 2000 caratteri circa; pertanto se l'utente non specifica il metodo di default viene usato Post.

Se la chiamata é andata a buon fine il json di ritorno viene mappato in un DTO specifico, a seconda del servizio scelto, e restitiuito direttamente.

```
public Task<T> CallApiAsync<T>(string uri, List<KeyValuePair<string, string>> content,
       HttpMethod method = null)
2
       var result = new HttpResponseMessage();
3
       if (method == null)
4
       {
          method = HttpMethod.Post;
       }
       if (_client.BaseAddress == null)
           _client.BaseAddress = new Uri(Localizations.BaseUrl);
11
       }
12
13
       return Task.Run(async () =>
14
           if (method == HttpMethod.Get)
              string query;
18
              using (var encodedContent = new FormUrlEncodedContent(content))
19
20
                  query = encodedContent.ReadAsStringAsync().Result;
21
              }
              result = await _client.GetAsync($"{uri}/?{query}");
          else if (method == HttpMethod.Delete)
           {
26
              string query;
              using (var encodedContent = new FormUrlEncodedContent(content))
              {
29
                  query = encodedContent.ReadAsStringAsync().Result;
30
              }
              result = await _client.DeleteAsync($"{uri}/?{query}");
          }
33
          else
34
           ₹
35
              var httpContent = new HttpRequestMessage(method, uri);
36
              if (content != null)
37
              {
38
                  httpContent.Content = new FormUrlEncodedContent(content.ToArray());
39
              }
40
              result = await _client.SendAsync(httpContent);
41
42
          string resultContent = await result.Content.ReadAsStringAsync();
43
          if (result.StatusCode == System.Net.HttpStatusCode.RequestUriTooLong)
44
```

```
{
45
               throw new ArgumentException(ErrorMessages.UriTooLong);
46
           }
           if
              (!result.IsSuccessStatusCode)
48
           {
49
               throw new Exception(resultContent);
50
           }
51
           return JsonConvert.DeserializeObject<T>(resultContent);
       });
   }
54
```

1.4 Testing

La maggiorparte dei servizi é stata testata tramite il tool XUnit; principalmente sono stati eseguiti test di validazione, data la natura della libreria, appoggiandosi ad una fixture comune per l'inizializzazione di ogni test.

É stato usato fin da subito il servizio online Travis di testing automatico, tramite il quale é stato possibile validare ogni rilascio facendo partire automaticamente i test con l'evento di push di git.

1.5 Documentazione

Infine le classi e i metodi piú rilevanti sono stati commentati tramite i tag XML specificati nella documentazione Microsoft e la documentazione in formato HTML é stata generata automaticamente tramite il tool Wyam.

1.6 Nuget

La dll generata dalla compilazione della libreria é stata infine documentata sotto il profilo delle dipendenze, generando il file .nuspec ed inclusa nel file .nupkg tramite l'apposito tool fornito da Nuget. Il tutto é stato caricato sul portale online Nuget ed é ora possibile includerlo in un progetto tramite il comando cli:

\$ dotnet add package SpazioDati.Dandelion

2 Calcolo della Relatedness

La seconda parte del progetto riguarda l'analisi di due metodi critici (readDump() e rel(int a, int b)) che servono a calcolare, dati gli identificativi di due entitá, il loro valore di correlazione (relatedness).

Per fare ció si ha a disposizione un dump dove sono salvati tutti i valori di correlazione per ogni coppia di entitá.

Il metodo readDump() serve a caricare in memoria i valori del dump sottoforma di "matrice" mentre il metodo $rel(int\ a,\ int\ b)$ serve a fare una ricerca nella struttura dati per poi ritornare il valore di correlazione fra le entitá con identificativi a e b.

Questo approccio attualmente funzionante é sicuramente molto efficiente in termini di prestazioni ma sicuramente molto oneroso in termini di memoria; pertanto il fine dell'analisi sarebbe quello di studiare/testare implementazioni alternative per ottimizzare le prestazioni e/o diminuire lo spreco di memoria.

2.1 Implementazione Iniziale

2.1.1 Il Dump

53676192

```
3 0.01

4 2

5 1.0

6 53676158 7 null

7 53676164 5 NnAwQT/ZvTVJ05MeSfcdIOrMrg1LBmUH

8 53676016 13 RDsFRUkyASRLFf3T

9 53675811 16 null
```

Leggendo le prime righe di un dump si nota un primo intero, denominato max_id , che indica il limite massimo che gli id delle entità possono assumere; dato che gli id delle entità non sono necessariamente sequenziali (fra due di essi potrei avere dei "buchi", degli intervalli in cui gli identificativi non sono associati ad alcuna entità) esiste una funzione map che mappa ("compatta") gli id delle entità su altri id univoci e sequenziali.

Il secondo intero, denominato nodesSize é il valore massimo che la funzione map puó assumere (limita il codominio della funzione).

Continuando a leggere il dump troviamo la relatedness minima considerata minRel (sotto la quale i valori di correlazione non vengono salvati) e altri valori di configurazione quali minIntersection e threshold che tuttavia non ci interessano particolarmente.

Da qui in poi inizia la definizione della matrice, ogni riga é composta da tre valori separati da uno spazio. Ad esempio se consideriamo la riga:

53676164 5 NnAwQT/ZvTVJ05MeSfcdI0rMrg1LBmUH

Il primo valore a=53676164 é l'id di un entitá, il secondo valore b=5=map(a) é il risultato della funzione map, infine il terzo valore c=NnAwQT/ZvTVJ05MeSfcdI0rMrg1LBmUH é una stringa di dimensione variabile oppure null; quest'ultima stringa, denominata postingList, é la codifica in Base64 di un array di byte che rappresenta un dizionario chiave valore.

In pratica una tupla $< chiave, \ valore >$ é codificata sull'array in gruppi di quattro byte; quindi partizionando l'array in gruppi da 4 byte ottengo, per ogni gruppo, la chiave (é un id "mappato") codificata sui primi tre byte ed il valore (la relatedness) definito sul quarto byte (il byte più a destra).

Questa codifica, pertanto, permette di gestire set di entitá con id "mappato" massimo $2^{3\cdot 8} \simeq 1, 7\cdot 10^7$ ed una scala di valori di correlazione appartenenti all'intervallo [0, 255].

Bisogna precisare che l'array di byte, denominato postingList, é ordinato per valore crescente delle chiavi e, dato che rel(a, b) = rel(b, a), si é deciso di salvare una sola volta il valore di correlazione imponendo la relazione: a > b.

Esempio

Nell'ipotesi in cui abbia una riga nel dump costituita dalla tupla (a, b, c) e volessi trovare relatedness(a, d) con map(a) > map(d).

Devo cercare nella $postingList\ c$, un gruppo di quattro byte in cui i primi tre byte corrispondano a map(d) ed il quarto byte sará il valore di correlazione voluto.

Ovviamente é possibile che c sia null (l'entitá con id a non ha nessuna correlazione con entitá i cui id "mappati" siano inferiori a b) oppure che nella posting non sia presente nessuna chiave corrispondente a map(d), in questi casi relatedness(a,d) = 0.

2.1.2 Il Codice Nativo

```
// STRUTTURA DATI
public class RelatednessMatrix {
   public int[] map;
   public byte[][] matrix;
   public float configuredMinRel;
   public int configuredMinIntersection;
   public float threshold;
   public static int EL_SIZE = 4;
```

```
}
9
   // CODICE UTILIZZO
   public float rel(int a, int b){
12
       if(a==b) return 1f;
14
       int nodeA = data.map[a];
       int nodeB = data.map[b];
16
       if (nodeA < 0 || nodeB < 0) return Of;</pre>
       int key = a>b ? nodeB : nodeA;
       byte[] array = a>b ? data.matrix[nodeA] : data.matrix[nodeB];
20
       if (array == null) return Of;
22
       int start = 0;
24
       int end = array.length - RelatednessMatrix.EL_SIZE;
       int pos = -1;
       while(pos == -1 && start <= end)</pre>
           int idx = ((start+end)/RelatednessMatrix.EL_SIZE)/2;
29
           int idx_value = ((array[idx*RelatednessMatrix.EL_SIZE] & 0xFF) << 16)</pre>
30
                      + ((array[idx*RelatednessMatrix.EL_SIZE+1] & 0xFF) << 8)
31
                      + ( array[idx*RelatednessMatrix.EL_SIZE+2] & 0xFF);
           if(idx_value == key) pos = idx;
35
               if(key > idx_value)
36
                  start = (idx + 1) * RelatednessMatrix.EL_SIZE;
               else
                  end = (idx - 1) * RelatednessMatrix.EL_SIZE;
           }
40
       }
       if(pos == -1) return Of;
43
       else
44
       {
45
           byte by = array[pos * RelatednessMatrix.EL_SIZE + 3];
           int byint = by + 128;
           float byteRel = byint / 255f;
           return data.configuredMinRel + byteRel * (1 - data.configuredMinRel);
49
   }
51
52
   // CODICE LETTURA
   public String END_OF_FILE = "" + '\0';
54
   public RelatednessMatrix readDump() throws Exception {
56
       URL url = getClass().getResource("dump.txt");
57
       File file = new File(url.getPath());
58
       BufferedReader fbr = new BufferedReader(new FileReader(file));
59
       String line = new String();
60
61
       try {
62
           int max_id = Integer.parseInt(fbr.readLine().trim()); // Trims all leading and
63
               trailing whitespace from this string
           int nodesSize = Integer.parseInt(fbr.readLine().trim());
           float minRel = Float.parseFloat(fbr.readLine().trim());
65
           int minIntersection = Integer.parseInt(fbr.readLine().trim());
66
           float threshold = Float.parseFloat(fbr.readLine().trim());
67
```

```
RelatednessMatrix data = new RelatednessMatrix();
           data.map = new int[max_id + 1];
70
           data.matrix = new byte[nodesSize][];
           data.configuredMinRel = minRel;
72
           data.configuredMinIntersection = minIntersection;
73
           data.threshold = threshold;
74
           int idx = 0;
76
           while ((line = fbr.readLine()) != null) {
               if (line.trim().equals(END_OF_FILE.trim())) {
                   break;
80
               }
81
               System.out.println(line);
82
               String[] splittedLine = line.split("\\s+"); // split(" ")
83
               if (splittedLine.length != 3) {
84
                   throw new Exception("Wrong format relatedness file for the line: " + line);
85
               }
87
               int wid = Integer.parseInt(splittedLine[0]);
88
               int node = Integer.parseInt(splittedLine[1]);
89
               data.map[wid] = node;
90
91
               if (splittedLine[2].equals("null")) {
                   data.matrix[node] = null;
93
               } else {
                   byte[] a = Base64.getDecoder().decode(splittedLine[2].toString());
95
                   data.matrix[node] = Base64.getDecoder().decode(splittedLine[2].toString());
96
               }
97
               idx++;
98
           }
99
100
           if (idx != nodesSize) {
               throw new Exception("Wrong format relatedness file the number of line do not
                   match with size of matrix");
           }
104
           return data;
        } catch (Exception e) {
106
           System.out.println(e.getMessage());
107
           return null;
108
        }
    }
```

Il metodo readDump() legge riga per riga il dump e restituisce un'istanza di RelatednessMatrix valorizzata. La funzione map viene salvata in un array di interi (data.map) mentre la matrice con le correlazioni viene salvata in un array bidimensionale di byte (data.matrix).

Il metodo $rel(int\ a,\ int\ b)$ invece calcola la funzione map sugli id a e b dopodiché fa una binary search sulla postingList ottenuta da matrix[map(a)] (assumendo map(a) > map(b)). Infine se la ricerca binaria ha successo il valore di relatedness ottenuto va riscalato su una scala di valori [0,255], convertito in float e riscalato nuovamente in base al valore di relatedness minima considerata (per escludere i valori al di sotto della relatedness minima, che non verranno mai usati).

```
byte by = array[pos * RelatednessMatrix.EL_SIZE + 3]; //considero il quarto byte
int byint = by + 128; //scalo di 128 valori per ottenere una relatedness in [0, 255]

//cast in float, escludendo i valori sotto la relatedness minima (configuredMinRel)
float byteRel = byint / 255f;
return data.configuredMinRel + byteRel * (1 - data.configuredMinRel);
```

2.1.3 Osservazioni

Considerato che il metodo readDump() viene chiamato una sola volta per allocare in memoria la struttura dati nel server, non ci sono particolari vincoli temporali per la sua esecuzione; per quanto riguarda il calcolo della relatedness invece l'algoritmo dev'essere il piú performane possibile.

Il codice nativo per il calcolo della relatedness costa $O(lg_2(n))$ (con n = nodeSize) per la ricerca binaria e O(1) per quanto riguarda le istruzioni precedenti e successive a quest'ultima.

Tolto un piccolo refact del codice che portebbe ad evitare tre moltiplicazione per ciclo nella binary search (non é necessario dividere e poi moltiplicare per RelatednessMatrix.EL_SIZE), non sembra si possono migliorare di molto le prestazioni dell'algoritmo senza cambiare struttura dati.

Una via percorribile per non "pagare" $O(lg_2(n))$ per la binary search potrebbe essere quella di allocare direttamente in memoria la matrice (in questo caso avrei un algoritmo di tempo costante O(1)). Questa soluzione peró andrebbe a peggiorare drasticamente lo spreco di memoria.

Stima della Memoria Allocata

Se consideriamo le prime righe del dump in questione possiamo stimare che mediamente il 57% degli id hanno postingList vuota (null) ed il restante 43% ha un numero medio di relatedness per id pari allo 0,000101% di nodeSize.

L'implementazione attuale alloca in memoria per l'array di interi map circa $nodeSize \cdot 4$ $Byte = 4^2$ $cdot10^6Byte = 16MB$ (arrotondando e senza tenere conto di fattori secondari come i 32 byte occupati da overhead e puntatore).

Per la matrice invece possiamo assumere che, su una macchina a 64 bit, avró il 57% di righe a null $(0.57 \cdot 4 \cdot 10^6 \cdot 8Byte = 18MB)$ ed il restante 43% di righe con mediamente 404 valori di relatedness codificati su 4 Byte $(0.43 \cdot 4 \cdot 10^6 \cdot 404 \cdot 4Byte = 2.8GB)$. Quindi in totale potremmo stimare che solo questo dump occupa quasi 3 GB (notare che esiste un dump per ogni lingua supportata da Dandelion e che questo é uno dei dump piú piccoli).

Se allocassimo in memoria una matrice di Byte completa di dimensioni nodeSize x nodeSize allocheremmo $(4 \cdot 10^6)^2 = 1.6 \cdot 10^1 3 Byte = 16 TB$ in memoria; bisogna considerare comunque che esistono strutture dati particolarmente ottimizzate per gestire array di grandi dimensioni ma con pochi valori (teniamo presente che più della metà dei valori nella matrice non saranno valorizzati) come SparseArray e SuperArrayList. Per quel che ho visto queste implementazioni si basano su liste ma con un'implementazione basata su liste difficilmente otterró un tempo di lookup di O(1) molto probabilmente avró un tempo di lookup $O(lg_2(noseSize))$ se non superiore. Pertanto a meno di uno spreco di memoria ulteriore difficilmente otterremo buoni risultati mantenendo come struttura dati di riferimento la matrice.

Una via praticabile potrebbe essere quella di vedere il problema al contrario, infatti ho solo 255 valori possibili di relatedness. Per minimizzare lo spreco di memoria Potrei pensare di allocare 255 array in cui storicizzo tutte le coppie di id che hanno la stessa relatedness. In questo caso peró non risparmio memoria, anzi peggioro, perché effettivamente non salvo piú volte l'informazione della relatedness ma le coppie di id sono interi codificati su 4 byte. Andrei ad occupare $0.43 \cdot 4 \cdot 10^6 \cdot 404 \cdot 2 \cdot 4 = 5.6GB$, occupo il doppio rispetto all'implementazione nativa, senza contare che le performance peggiorerebbero.

Un'altra via potrebbe essere adottare un hash map che mappi gli id concatenati sul valore della relatedness, se la funzione fosse ben ottimizzata otterrei in O(1) la relatedness e se occupasse meno memoria della matrice iniziale sarebbe un ottima soluzione al problema.

Allegato A Titolo primo allegato

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosque ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

A.1 Titolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

A.1.1 Sottotitolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosque ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

Allegato B Titolo secondo allegato

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosque ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

B.1 Titolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosque ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

B.1.1 Sottotitolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosque ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.