Semester project report

Department of Telecommunications and Media Informatics

Author:	**Márton Torner**
Neptun:	**OZFKFP**
Specialization:	**Infocommunication (HIT)**
E-mail address:	**torner.marton@gmail.com**
Supervisor:	**Bálint Gyires-Tóth, PhD**
E-mail addres:	**toth.b@tmit.bme.hu**

Title: Financial time series modeling and forecasting with deep neural networks

Task

The student's task is to collect limit order book data from a chosen crypto exchange, analyze and process the gathered records and to develop a neural network based solution to predict the direction of the price movement of the asset-pairs.

In this work level 100 data (limit order book, depth: 100) is used from Kraken crypto exchange. The live order flow is queried with the help of the provided API and saved on a local storage. The aim is to create a Deep Learning powered system which can predict the price movement direction from the given history. The first approach is a linear regression based solution and after the successful tests I try to use a deep neural network as the model.

In the future this work can be used as a base for creating a complex system for algorithmic trading.

2018/19 II. semester

1 Theory and previous works

1.1 Introduction

The computerization of financial markets and the availability of the electronic records of different stock exchanges like transactions, order flow and limit order books provide us a great opportunity to analyze and model the dynamics of these markets. Over the last years a huge number of studies have been done on this field which can serve as a good base for further studies.

With the computerization also came the rapid acceleration of the trading on these financial markets and big companies started to use algorithmic solutions to steal a march on their rivals and gain some higher profits.

The quick development that the field of Deep Learning had in the last decade and the impressive number of successful researches conducted on the topic of time series forecasting opened the door to change the traditional mathematical algorithms with a neural network based solutions. Price forecasting is just a step leading to more complex systems which use the predictions as inputs to deliver a trading strategy with the help of machine learning (reinforcement learning).

These solutions could change the role of financial specialists in the future.

1.2 Theoretical summary

1.2.1 Price formation mechanism

The market price per share of stock ("share price") is the amount of money an investor is willing to pay to own a company's single stock. The 'price formation mechanism' - given in Equation 1 - is a map which tries to represent the relation between market price and different variables affecting it, e.g. price history, order flow, news about the companies or the market etc.

$$Price(t + \Delta t) = F(Pricehistory(0...t), OrderFlow(0...t), OtherInfo) = F(X_t, \epsilon_t) \quad (1)$$

X_t: set of state variables (e.g. order flow, volatility, price, etc.)
ϵ_t: random noise (represents the arrival of new, unknown information)
Δt: time resoution

It is shown that this relation is universal (not asset-specific) and stationary (stable across time, even at out-of-sample predictions) [1]. These two features give us the opportunity to use all collected data points to design and train a universal model which can make predictions for all asset-pairs valid even in the future.

1.2.2 Limit Order Book

In this work limit order book data is used as the input so this section tries to provide a short overview what it is and how it works. On stock exchanges we want to buy shares or sell the ones we hold and to accomplish this we place orders of which the LOB consists. There are more special types of orders but in this work only limit orders are used. They remain in the LOB until being executed or cancelled.

The LOB has two sides, the ask and the bid side, shown on Figure 1. The ask side represents the sell orders (willing to sell the given amount at the given price) and the bid side the buy orders (willing to buy the given amount at the given price).

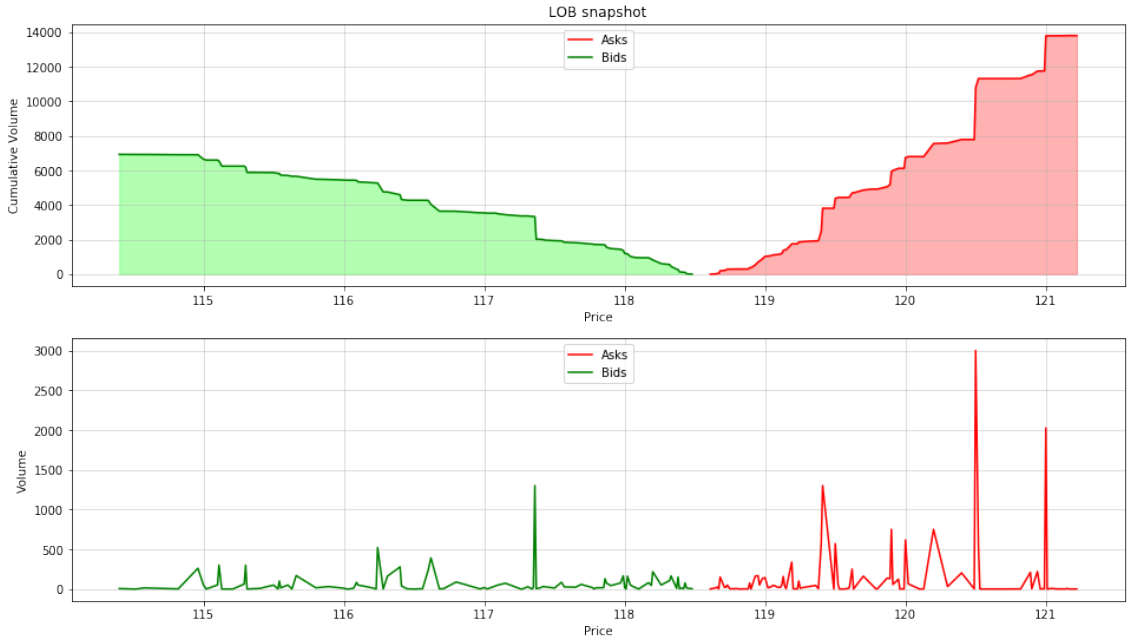

Figure 1: Limit Order Book snapshot (ETH_EUR). Bottom: Volume at each price level. Top: Cumulative volume at price levels.

A tick is the measure of the minimum upward or downward movement in a price of a security this means the price levels in the LOB must be the multiples of this value (so order prices can not differ less than this). The best ask price is the lowest sell order and the best bid price is the highest buy order. These are the prices at which the given number of shares can be bought or sold. The difference between best ask and bid price is called the spread and their average is the mid-price. The weighted average mid-price (WAMP) and volume weighted average price (VWAP) can be calculated as given in Equation 2 and Equation 3. The depth of the book is used in this paper as the number of best orders on each side (so depth=100 means the top 100 lowest priced ask- and the top 100 highest priced bid orders).

$$WAMP = \frac{p^{bid}_{best} * v^{ask}_{best} + p^{ask}_{best} * v^{bid}_{best}}{v^{bid}_{best} + v^{ask}_{best}} \tag{2}$$

$$VWAP = \frac{(\sum_{i=1}^{depth}(p^b_i * v^b_i)) * v^a + (\sum_{i=1}^{depth}(p^a_i * v^a_i)) * v^b}{v^a + v^b} \tag{3}$$

p: price
v: volume (without lower index = sum over side)
b: bid
a: ask

Orders are submitted and cancelled continuously which means that the updates arrive in a very high frequency (millisecond) and it is also easy to see that consuming the order flow of stock exchanges leads to Terabytes of data. At crypto exchanges order manipulation is a big problem to be faced at filtering the collected data. So, to process a live order book we have to deal with various problems [TODO link ide a crawler szekcióhoz].

1.2.3 Deep Learning

Deep Learning models are deep neural networks (an example can be seen on Figure 2) trained on large datasets to uncover complex non-linear relations between the (high-dimensional) inputs and outputs. This can be simplified as it represents a functional relation like y = f(x) where y is the output and x is a (high-dimensional) input vector.

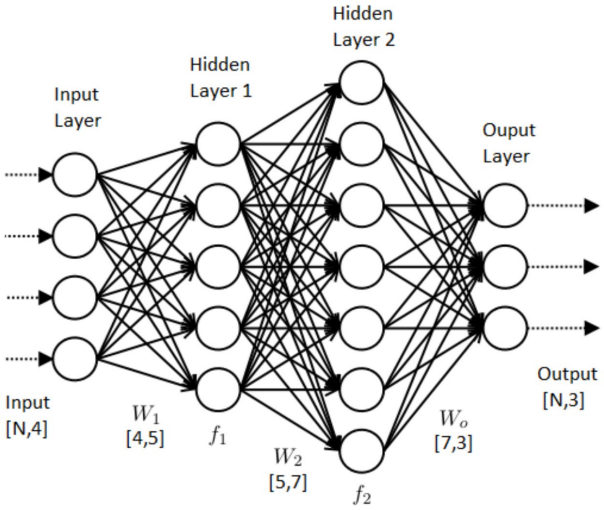

Figure 2: Simple multilayer perceptron.

The network iterates the input data through its hidden layer(s) (basically weighted sums are calculated followed by so called 'activation' functions). In supervised approaches the weights are modified in the backpropagation process in which a regularized cost function – reflecting the discrepancy between the inputs and the desired output – is tried to be optimized. Depending on the input and output dimensionality and the design of the network it can have millions of parameters which makes these calculations computationally expensive. This problem made the development of this field a standstill until the rapid spread and improvement of GPUs started providing a huge advancement in calculations.

1.2.4 Convolutional Networks

A class of deep neural networks are convolutional neural networks (CNN). The main design is the same as any other networks (input, hidden layers, output). General CNNs use a set of convolutional layers followed by fully-connected layers (all neurons in a layer are connected to every neuron in the preceding and following layers as seen on Figure 2) as hidden layers. Convolutional layers are used for representing small or high features in the data with applying a convolutional operation to their input. It uses filters which are parts of the data with the same or less dimension than the input. Using k filters k feature maps are generated.

Attention is a mechanism which can be added to reach better results with the cost of more computational expenses [2]. At a high-level attention can be described as giving the network the ability to learn which parts are more important to focus on. It is commonly and successfully used at natural language processing tasks.

1.2.5 Generative Adversarial Networks

Generative adversarial networks (GAN) are deep neural architectures comprised of a generator and a discriminator network (Figure 3), introduced by Goodfellow et al. [3]. These two networks

contest against each other in the following way. The aim of the generator is to produce data indistinguishable of the real data and the discriminator tries to classify whether its input was real data or fake, generated by the opposing network. The losses of the networks are propagated back combined making both networks' reward depending on both of their performance. This means if the generator lacks competence the discriminator's work will be easy thus the system will be incompatible of solving the problem. The design and training of GANs is a complex task.

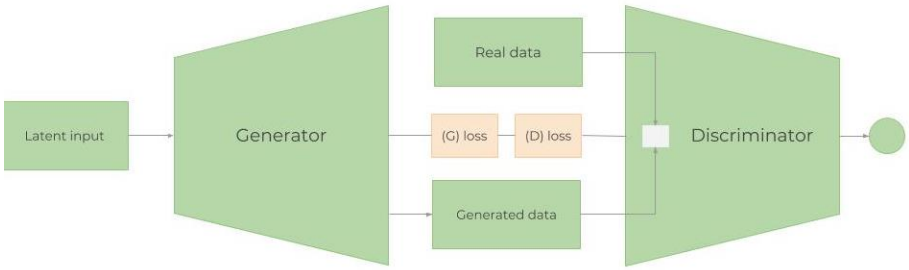

Figure 3: GAN architecture example.

1.3 Starting point, previous works on this project

At the start of the project I was no preceding work given and did not make any research on this topic before.

I started learning the theory of Deep Learning as part of the university course VITMAV45 and gained a well based knowledge on the subject to start with.

2 Work on project

2.1 Data

The project's first task was to find or collect a database to work with. On financial field the big and high frequency raw datasets which are precise enough can cost a lot of money (3000 USD+). The was no funding available so I collected my own raw data. It is from Kraken (www.kraken.com) which is a US based cryptocurrency exchange. Kraken gives the opportunity to access live order book (and of course other) data through a websocket based API.

In the training and evaluation process only ETH_EUR [1] snapshots are used but for further research on universality and stationarity [1] I save 25 asset-pair's data in aggregate. For each 1 snapshot is querried every minute through REST API calls and the real-time order flow is recorded through the previosly mentioned Websocket API.

The saved data is stored in CSV file format, Table 1 and Table 2 show the structure. Every asset-pair is saved to a separate folder and at the end of each day all buffer files (updates and snapshots) are compressed and saved into a separate file this way saving space and ensuring that the unexpected corruption of a file causes minimal data loss. The saving method also makes paralell datapreparation on multiple CPUs possible. The dataset contains entries from 15th March 2019.

Field	Description
L1 Ask Price	The price of the ask order at 1st level (best ask)
L1 Ask Volume	The volume of the ask order at 1st level
L1 Ask Timestamp	The last modification time of the ask order at 1st level
...	
L100 Ask Price	The price of the ask order at 100th level
L100 Ask Volume	The volume of the ask order at 100th level
L100 Ask Timestamp	The last modification time of the ask order at 100th level
L1 Bid Price	The price of the bid order at 1st level (best bid)
L1 Bid Volume	The volume of the bid order at 1st level
L1 Bid Timestamp	The last modification time of the bid order at 1st level
...	
L100 Bid Price	The price of the bid order at 100th level
L100 Bid Volume	The volume of the bid order at 100th level
L100 Bid Timestamp	The last modification time of the bid order at 100th level
Timestamp	The timestamp of the snapshot

Table 1: One snapshot (one row in snapshot files).

Field	Description
AskOrBid	The update type: 0 = ask, 1 = bid
Price	The price of the order
Volume	The volume of the order (0 means that the level is deleted)
Timestamp	The last modification time

Table 2: One update (one row in update files).

[1] Ethereum - Euro: Ethereum is an open-source, public, blockchain-based distributed computing platform and operating system featuring smart contract (scripting) functionality. - *Wikipedia*

2.2 A munkám ismertetése logikus fejezetekre tagoltan

<Én magam (nem a társam) a félév során következőket olvastam el / programoztam / készítettem el / teszteltem / dokumentáltam / néztem át / tanultam meg, stb. Tételes leírása és felsorolása mindannak, ami a félév során történt, alátámasztandó azon állításom a konzulens/tárgyfelelős felé, hogy összességében mindent beleértve tényleg dolgoztam a TVSZ szerint kreditenként 30 órát, azaz a heti 2 kontakt órás tárgy esetében min. $2,5 * 30 = 75$ munkaórát, illetve a heti 6 kontakt órás tárgy esetében min. $8 * 30 = 240$ munkaórát...>

Ebben a részben a hallgató az általa elvégzett munkát mutatja be. Hangsúlyosan a saját munka bemutatása a cél, hiszen a hallgató ezzel igazolja a témavezető és a tárgyfelelős irányába, hogy – folyamatosan fejlődve és egyre több és jobb munkát végezve – a szakdolgozatát/diplomadolgozatát képes lesz megírni. A beszámoló nem munkanapló, nem arra vagyunk kíváncsiak, hogy mit mikor csinált a hallgató és mennyi időt töltött vele, hanem egy eredmény-centrikus beszámolót szeretnénk olvasni. De itt is fontos tudni, hogy megosztott task esetén ki-mit csinált, mekkora részt vállalt.

Az egész beszámoló elkészítésénél törekedni kell a magyar nyelv szabályainak követésére és a műszaki dokumentáció/tudományos közlemény írásával kapcsolatosan kialakult közmegegyezés szerinti formai követelmények betartására. (Tehát nem kell többes számként hivatkozni saját magunkra, kerülni kell a furcsa megfogalmazást, passzív és egyéb kifacsart mondatszerkezeteket. Az egy szót határozatlan névelőként történő használatakor ne írjuk ki számként.)

A beszámoló természetesen nem csak szöveget tartalmazhat, hanem képleteket, táblázatokat, ábrákat és még sok minden mást. Ezek kapcsán az alábbi elvek irányadók:

- Az ábráknak, képeknek és táblázatoknak mindig van számuk és címük. (A cím nem ennyi: „1. ábra”, hanem azt írd le, ami látható rajta.)
- Az ábrákra, a képekre és a táblázatokra a szövegben hivatkozni kell, és a szövegben elemezni kell azokat. Például
- Az ábrák, képek és táblázatok mérete a szükségesnek megfelelő legyen: elég nagy ahhoz, hogy kinyomtatva is olvasható és értelmezhető legyen, de nem nagyobb annál, mint amit szerepe indokol.
- A grafikonoknak a tengelyeken legyenek feliratai és ha releváns, a mértékegység is.
- A képletek esetében nem minden képletre történik hivatkozás, de ahol igen, ott a képletet a műszaki irodalomban jellemző módon a sor végére tett kerek zárójelben lévő számmal jelöljük meg. A képleteket ne képként illeszd be a szövegbe.
- Kódrészleteket, ha nem relevánsak, ne illeszd be képként, főleg ne rossz minőségben. Nyugodtan teheted függelékbe és hivatkozd be a szövegben, mint a képeket, például: Az 1. számú függelékben található az adatbeolvasó kód, melyet C++ nyelven készítettem el.

Az írásbeli beszámolót a témavezető és a tárgyfelelős is értékeli. A tárgyfelelősi értékelés szempontjai az alábbiak:

1. Megfelel-e az elvégzett munka a félév elején kiadott tasknak?
2. Megfele-e a beszámoló a formai követelményeknek? Ezen belül:
    - a. Megfelelő-e az elméleti bevezető és az irodalomjegyzék?
    - b. Egyértelmű-e, hogy mi volt a hallgató saját munkája?
    - c. Megfelelő-e a dokumentum technikai színvonala?

Az írásbeli beszámoló beadásának napja a szóbeli beszámolóhoz képest (munkanapban)	A „b” faktor értéke
-4. munkanap	0.04
-3. munkanap	0.09
-2. munkanap	0.20
-1. munkanap	0.30

Table 3: Az írásbeli beszámoló késedelmes beadásával kapcsolatos hanyagsági faktor értéke

Ezen kívül a tárgyfelelős veszi figyelembe az értékelés során kialakult félévi jegyre vonatkoztatva az ún. „hanyagsági faktor” értékét, amelyet (2) szerint állapítunk meg:

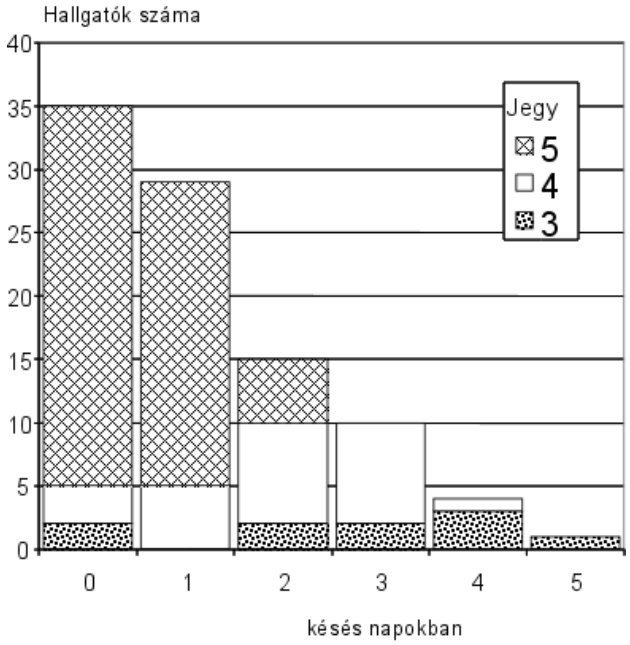

Figure 4: Hallgatók érdemjegyeinek eloszlása az írásbeli beszámoló késése függvényében

Az (1)-ben szereplő a szám a munkaterv beadásában történt késedelemre, míg a b szám az írásbeli beszámoló beadásában történt késedelemre vonatkozik. Utóbbi értékeiről

A beszámoló értékeléséről részletesebben írunk [6]-ban.

A beszámolóban bizonyára szerepelni fognak rövidítések. Ezeket a rövidítéseket, betűszavakat néhány, az infokommunikáció területén nagyon ismert és gyakran használt kifejezéstől (például IP, TCP, GPRS, UMTS) eltekintve ki kell fejteni logikusan az első használat alkalmával (például így: „A GPS (Generalized Processor Sharing) egy ideális folyadékmodellen alapuló csomagütemező eljárás.”).

A beszámoló készítése során előfordulhat, hogy a hallgató úgy érzi, hogy alfejezetekkel tagolva jobban olvasható és érthető lenne a beszámoló. Ennek akadálya nincs, de érdemes arra figyelni, hogy a túlzott tagolás sem tesz jót egy írásműnek, illetve hogy a címsorokban a rövidítések és a hivatkozások használata tilos. Tartalomjegyzéket készíteni nem szükséges a beszámolóhoz, de nem is tilos, kivéve azt az esete, amikor nyilvánvalóan terjedelemnövelési célokat szolgál.

A beszámoló terjedelme tárgyanként változhat. Általános szabály, hogy 1 hüvelyknél nagyobb margókat ne használjunk. A szöveg legyen egyszeres sortávú, sorkizárt és 12 pontos betűméretű. A bekezdések kezdődjenek behúzással a minta szerint.

2.3 Összefoglalás

A félévi munka során elért új eredmények ismételt, vázlatos, tömör Ebben a részben az adott félévre vonatkozó, az *Önálló laboratórium tárgy keretében elvégzett munka során* elért **új** eredmények ismételt, vázlatos, **tömör** összefoglalását várjuk, lehetőleg nem felsorolásként. Itt még egyszer ki lehet térni a leglényegesebb eredményekre, valamint a félév során felmerülő nehézségekre, de meg lehet említeni a továbbfejlesztési irányokat, lehetőségeket is.

Ezt a részt tagolható a következő pontok megválaszolásával:

- Mi volt az **aktuális kérdés**, probléma, amivel a félév során foglalkoztál?
- Mi a dolgozat **célja**, miért érdekes egyáltalán ezzel a problémával foglalkozni?
- Milyen **módszereket** használtál a probléma megoldása érdekében?
- Mik a legfontosabb **eredmények**?
- Milyen **következtetéseket** lehet levonni?

Ha valaki elolvassa ezt a részt, képet kell kapnia az egész dolgozatról. Ne legyen az absztrakt szó szerinti ismétlése.

Fontos, hogy az itt megadott sablontól el lehet térni, használata nem kötelező, csak segítséget jelenthet, viszont a fedőlap lehetőleg maradjon ugyanez és tartalmilag egyezzen meg a sablon irányelveivel. A beszámoló felépítésében nem érdemes eltérni a *Bevezető – Féléves munka és eredmények bemutatása – Összefoglaló* hármastól.

3 References

[1] Justin Sirignano and Rama Cont (2018). *Universal features of price formation in financial markets: perspectives from Deep Learning*

[2] https://medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio (2014). *Generative Adversarial Nets*

[4] Umberto Eco, *Hogyan írjunk szakdolgozatot?*, Kairosz Kiadó, 2000, ISBN: 9639137537.

[5] Esterházy Péter, *Termelési-regény (Kisssregény),* Magvető Könyvkiadó, 2004, ISBN: 9631423948.

[6] *Tájékoztató a Műszaki Informatika Szak önálló laboratórium tantárgyainak 2008/9. tanév I. félévi lezárásáról a BME TMIT-en (VITMA367, VITMA380, VITT4353, VITT4330),* http://inflab.tmit.bme.hu/08o/lezar.shtml, szerk.: Németh Felicián, 2008. november 5.

[7] Wikipedia contributors, *Wikipedia:Academic use*, Wikipedia, The Free Encyclopedia, 2011 Nov 11. Available from: http://en.wikipedia.org/w/index.php?title=Wikipedia:Academic_use&oldid=460041928

Itt jegyezném meg, hogy a tanulmányozott irodalmat hivatkozni kell a szövegben. Szükség esetén többször is. Az irodalomjegyzék célja (lásd fejezetet) ugyanis kettős[2]:

1. Az olvasó tájékoztatása, hogy a dokumentumban ki nem fejtett dolgoknak, a tudottnak vélt ismereteknek hol lehet bővebben utánanézni, így ott kell meghivatkozni az irodalmat [4, 5], ahová az irodalom kapcsolódik.

2. Megmutatni a tárgyfelelosnek/konzulesnek az elolvasott irodalom mennyiségét

Javasoljuk, hogy a hallgatók tanulmányozzák, hogyan néznek ki a hivatkozások a villamosmérnöki/informatikai szakma vezető szakmai folyóirataiban megjelenő cikkekben. Ebben a témavezető is biztosan tud segíteni. A hivatkozás teljességére és egyértelműségére tessék ügyelni. Például, ha egy könyvnek több, eltérő kiadása is van, akkor azt is meg kell jelölni, hogy melyik kiadásra hivatkozunk. A webes hivatkozások problémásak szoktak lenni, de manapság egyre több az olyan dokumentum, ami csak weben lelhető fel, ezért használatuk nem zárható ki. Itt is törekedni kell azonban a pontosságra és a visszakereshetőségre. A weben található dokumentumoknak is van címe, szerzője, illetve érdemes megadni a letöltés/olvasás időpontját is, hiszen ezek a dokumentumok idővel megváltozhatnak.

A wikipédiás hivatkozások használata nem javasolt, mert a wikipedia másodlagos forrás. Tájékozodjuk a wikipédián, de aztán olvassuk el az adott oldalhoz megadott hivatkozásokat is. A wikipedián külön szócikk foglalkozik azzal, hogy miért nem szerencsés tudományos munkákban a wikipédiára hivatkozni [7].

Nem publikus dokumentumok hivatkozása nem javasolt és csak kivételes helyzetben elfogadható!

[2] Akárcsak ennek a fejezet hivatkozásnak, ami a `\aref babel` parancsot demonstrálja

3.1 A csatlakozó dokumentumok jegyzéke

<A munka ezen beszámolóba be nem fért eredményeinek (például a forrás fájlok, mindenképpen csatolni akart forráskód részlet, felhasználói leírások, programozói leírások (API), stb.) megnevezése, fellelhetőségi helyének pontos definíciója, mely alapján a az erőforrás előkereshető – értelemszerűen nem nyilvános dokumentumok hivatkozása nem elfogadható.>